Feminist Philosophy Quarterly

Volume 8 | Issue 3/4 Article 1

2022

The Bias Dilemma: The Ethics of Algorithmic Bias in Natural-Language Processing

Oisín Deery
York University and Macquarie University
oisin@oisindeery.com

Katherine Bailey
Shopify Inc.
katherine@katbailey.net

Recommended Citation
Deery, Oisín, and Katherine Bailey. 2022. "The Bias Dilemma: The Ethics of Algorithmic Bias in Natural-Language Processing." *Feminist Philosophy Quarterly* 8 (3/4). Article 1.

The Bias Dilemma:
The Ethics of Algorithmic Bias in Natural-Language Processing

Oisín Deery and Katherine Bailey*

Abstract

Addressing biases in natural-language processing (NLP) systems presents an underappreciated ethical dilemma, which we think underlies recent debates about bias in NLP models. In brief, even if we could eliminate bias from language models or their outputs, we would thereby often withhold descriptively or ethically useful information, despite avoiding perpetuating or amplifying bias. Yet if we do *not* debias, we *can* perpetuate or amplify bias, even if we retain relevant descriptively or ethically useful information. Understanding this dilemma provides for a useful way of rethinking the ethics of algorithmic bias in NLP.

Keywords: artificial intelligence, algorithms, bias

1. Introduction

Bias in statistical models or algorithms can result in moral harms. A model is an abstract representation of a phenomenon, such as how words relate in a natural language. We prefer the term "model" to "algorithm," since strictly speaking "algorithm" refers only to the actual sequence of steps carried out during the training of a model or at the time it issues predictions. However, everyday use of "algorithm" generally picks out the already-trained model, which implicitly includes the data on which that model was trained. If the data are biased, then there are biased algorithms.[1] Yet for simplicity's sake, we will mostly stick with "model."

Reliance on biased models can exacerbate marginalization of vulnerable groups. For instance, the ethical issues related to statistical models in hiring have been widely discussed (e.g., Hu and Chen 2017). Problems can arise even when models are deployed to *reduce* the influence of bias. Thus, recidivism models aim to reduce the

* Oisín Deery, Departments of Philosophy at Macquarie University, Sydney, Australia & York University, Toronto, Canada (oisin@oisindeery.com)
Katherine Bailey, Shopify Inc., Toronto, Canada (katherine@katbailey.net)

[1] Algorithms can be biased for many reasons (see sections 6.3 and 8). We focus on bias deriving from data.

influence of bias in decision-making when sentencing offenders yet often serve only to reproduce or disguise bias (Kehl, Guo, and Kessler 2017).[2]

More generally, numerous ethical questions arise in relation to machine learning (ML) and artificial intelligence (AI), most of which we leave aside. Instead, our focus is on bias in the data on which ML models are trained, and especially language models. The central issue here is that the corpora on which such models are trained typically contain biases—for example, sexist or racist biases. That is because the corpora comprise examples of how people actually use language, and people use language in sexist and racist ways.[3] Thus, the outputs of models trained on such data reflect biases (see, e.g., the papers surveyed by Blodgett et al. 2020). We must ask whether (1) it is practically speaking *possible* to debias language models and (2) whether we *should* do so, if possible.

In relation to (2), a central issue is whether there might be practical or ethical value to *not* debiasing models. Some maintain—as we do—that not debiasing can be valuable, since models often reflect descriptively accurate information about how language is actually used (see Goldberg 2021). By contrast, others claim that not debiasing runs the risk of perpetuating and amplifying biases (Bolukbasi et al. 2016). Some argue that existing models do not actually provide descriptively accurate information at all, since the corpora on which the models are trained do not reflect marginalized voices (Bender, Gebru et al. 2021).

These debates do not deploy the descriptive-normative distinction that we will outline, yet this distinction helps to clarify the points of contention in such debates. Furthermore, even if we *had* descriptively accurate models—for example, models that do not exclude marginalized voices—there might be practical and ethical value to *not* debiasing, especially when we can control for the negative effects of bias at the output stage rather than by debiasing the model itself.

Once we recognize that there might be an ethical value to not debiasing, we encounter an ethical dilemma. Either we do not debias, in which case we retain

[2] Statistical recidivism models aim at reducing the influence of biases (explicit or implicit) of judges in sentencing. Yet the data these models use are typically obtained from questionnaires that criminals are asked to complete. These data often include details about a criminal's upbringing, family, and social connections. Such data ought to be irrelevant to a criminal's sentencing. Nevertheless, since the data are used to generate a "recidivism score" for criminals, which is used in sentencing, the data influence sentencing in a way that it should not. For example, higher scores often result in longer sentences. More perniciously, a longer sentence will often subsequently result in a higher score on the recidivism scale.

[3] Data can also be biased due to bad sampling. We do not focus on such bias (but see sections 6.3 and 8).

descriptive accuracy and potentially ethically useful information yet run the risk of perpetuating or amplifying bias; or instead we debias, thereby avoiding perpetuating bias yet losing descriptive accuracy and ethically useful information. Our aim is to draw attention to this dilemma and to show how it can help in rethinking the ethics of algorithmic bias in NLP systems.

In section 2, we specify the type of bias we will focus on. In section 3, we outline the distinction between descriptive and normative correctness as often competing virtues of NLP models, by analogy with how biases rooted in statistical regularities are often thought to offer support for beliefs even when these beliefs are morally problematic. In section 4, we illustrate our distinction by focusing on a particular variety of NLP model called word-embeddings. In section 5, we discuss practical problems with debiasing NLP models. These problems lend support to our claim that it is often preferable to control for negative effects at the output stage rather than try to debias the models themselves. In section 6, we consider varying degrees of scope for debiasing NLP systems' outputs. In section 7, we explain the ethical dilemma. In section 8, we discuss how our view helps to rethink the ethics of algorithmic bias in NLP.

2. How (and How Not) to Control for Morally Relevant Bias

Both human cognition and artificial systems can exhibit morally relevant bias. In humans, the relevant bias is typically a negative evaluative tendency regarding other people based on their apparent membership in a socially salient category or group (Brownstein 2018, 43, 126, 172; Brownstein and Saul 2016a, 2016b). Such biases are often *explicit*, consisting of mental attitudes (like *belief*) that an agent has and consciously endorses. For example, someone is explicitly sexist if they openly admit to being sexist and approve of these attitudes. By contrast, *implicit* biases are unconscious attitudes that can be difficult to inhibit, seem insensitive to an agent's explicit attitudes, and are often in conflict with an agent's explicitly held attitudes (see Levy 2017 for a review of the empirical literature). For example, even someone who explicitly condemns sexism might still have implicit sexist attitudes that they are unaware of and that are difficult to inhibit.

When people *act* on explicit biases, we hold them morally responsible for doing so, and as a result we blame them for their behavior (see, e.g., Brownstein and Saul 2016b). It is more difficult to know whether we can blame or hold people responsible for actions caused by *implicit* biases, since it is unclear whether people have sufficient control over how their biases influence behavior for them to be responsible for these behaviors (Levy 2017).[4] Even so, it is plausible to think that we

[4] The view that control is *not* needed for responsibility is a minority position (e.g., Smith 2008). It is more widely held that responsibility of the sort that might justify

need, at least, to control for the negative influences of implicit bias when expressed in behavior.[5] There can be structural ways of doing so. For example, the biased hiring of male lead violinists in orchestras (whether explicit or implicit) can be controlled for by having musicians audition behind a screen so that the hiring panel cannot see the violinist's gender (see, e.g., Goldin and Rouse 2000).

Bias in the relevant sense is something that it can be ethically problematic for a system to have, and it is especially bad when a system's outputs have negative ethical effects by causing harms. Human cognition exhibits implicit bias, and various ML systems, including NLP systems, exhibit a close analog of such bias.[6] In each case, the relevant system learns or acquires its biases from statistical regularities it identifies in the world.[7] We humans learn biases from other members of our communities.[8] For ML systems, biases are often learned from training data, which for NLP systems are language corpora that we produce. Since we produce the data containing the biases, the biases the NLP systems acquire they learn from us.[9]

To preview one of our central claims: It can be difficult to debias human cognition, especially for implicit biases, and we think it is also difficult to debias ML

---

blaming or praising *does* require control. The empirical literature suggests that we *lack* such control over how implicit biases affect our behavior (Levy 2017).

[5] We make no claim about whether agents *can* be responsible for actions caused by their implicit attitudes—although one of us defends this view in print (Deery 2021). Here, we claim only that even when behaviors are caused by implicit attitudes, we need to do *something* to control for their negative effects.

[6] The relevant biases for ML systems are analogous to implicit biases because such biases are not explicitly represented in models but instead take the form of implicit distributed patterns—e.g., between vectors in a model. We expand on this analogy between human cognition and ML systems in section 3.

[7] Some human cognitive or perceptual biases may be innate. They are not our focus here.

[8] In cases where a population uses words in certain ways and people acquire biases from exposure to that usage, the exposure might be mediated by how the media portrays such language use. Even so, we acquire biases from what we are exposed to (the world), which here would include how the media portrays language use.

[9] Might we have more control over the inputs to ML systems than we do for ourselves? If so, we could control for biases at the *input stage*. Yet the corpora on which ML systems are trained are vast, and to eradicate biases from them would be an impractical task. That is why those who seek to eliminate bias from models typically focus on the already-trained models (see section 5 for details). Exceptions include badly sampled data and algorithms that are biased for other reasons (see sections 6.3 and 8 below).

systems, including NLP systems. For humans, it is often easiest to control for the negative effects of bias at the output stage. As noted already, auditioning musicians can perform behind a screen, thereby defusing the possibility of implicit biases' influencing hiring. As a general matter, it seems that the negative effects of bias are often best controlled for when we acknowledge bias in a system and control for its negative influences at the output stage rather than by debiasing the system itself.[10] The possibility we explore in this paper is whether we might do something similar for NLP systems. Moreover, we claim that there can be an ethical value to *not* debiasing ML systems or their outputs, and we maintain that the question of whether we *should* debias raises an ethical dilemma.

3. Descriptive vs. Normative Accuracy

To explain our claim that there can be an ethical value to not debiasing models or outputs, we rely on a particular way of distinguishing *descriptive* accuracy from *normative* correctness.[11] We explain our distinction by analogy with how implicit biases in human cognition can offer support for beliefs even when these beliefs are morally problematic. In this way, people's *epistemic* reasons can support biased beliefs even when their *ethical* reasons cut against these beliefs. In relation to racist bias, Rima Basu outlines this epistemic/ethical conflict as follows:

> We may find ourselves facing the following conflict: what if the evidence . . . supports something we morally shouldn't believe? For example, it is morally wrong to assume, solely on the basis of someone's skin color, that they're a staff member. But, what if you're in a context where, because of historical patterns of discrimination, someone's skin color is a very good indicator that they're a staff member? When this sort of normative conflict looms, a conflict between moral considerations on the one hand and what you epistemically ought to believe given the evidence on the other, what should we do? It might be unfair to assume that they're a staff member, but to ignore the evidence would mean risking inaccurate beliefs. Some . . . have suggested that we simply face a tragic irresolvable dilemma. (Basu 2020, 191; cf. Gendler 2011; Puddifoot 2017)

[10] If the bias comes from badly sampled data, that is a different matter.

[11] We introduce the normative/descriptive distinction at some length since we have found it is unfamiliar to nonphilosophers working in AI and ML. Yet these researchers find it useful when explained.

In other words, descriptively accurate information about statistical regularities can support beliefs that are morally problematic or normatively incorrect, in which case we face a choice about whether to prioritize the descriptive or instead the normative. Likewise, we think that there are two general ways in which ML systems or humans can be right or wrong—descriptively and normatively—and that a dilemma similar to the one Basu describes can arise even for ML systems.

To explain, a visual representation of a scene gets things descriptively right to the extent that it accurately depicts visual information about a scene as presented and wrong to the extent that it does not. For example, if you photograph me robbing a store, your camera gets things descriptively right insofar as it accurately depicts my face and wrong insofar as it does not—for example, because its exposure settings are off. Most tasks a camera is designed to perform require for success that at least some threshold of descriptive accuracy (in our sense) be achieved in the camera's outputs. Likewise, an audio recording of my boss making sexist or racist remarks is descriptively accurate to the extent that it records what he said and descriptively wrong to the extent that it does not, perhaps due to electromagnetic interference from another device.

By contrast, we expect people—as moral agents—to get things normatively right in a moral sense. It is normatively correct for me to pay for my groceries instead of stealing them (all else being equal). When I rob a store, by contrast, I get things normatively incorrect by failing to do what I ought to do (pay for my groceries) and by doing what I should not (stealing them). Similarly, my boss behaves normatively correctly by *not* making sexist or racist remarks.

We do not expect systems like cameras or audio devices to get things normatively right. Consider an imaginary camera, the iBlur, which blurs out the morally objectionable parts of a visual scene.[12] As a result, the iBlur blurs out the image of me robbing the store, since robbing a store is a morally objectionable action (all else being equal). Not only do we *not expect* a camera to work like this but we *expect it not* to work this way, since such an imposition of normative correctness undermines descriptive accuracy, which the functions of a camera require.[13] Additionally, the iBlur would fail to provide us with ethically useful information, since

[12] We model our iBlur after the Arkangel system in the TV show *Black Mirror*.

[13] Cameras serve various functions, including the attainment of artistic or imaginative goals. Likewise, security cameras might be used to intimidate people or to shed light on the activities of select people in select places in a way that is morally problematic. Yet the tasks cameras perform still require that cameras accurately depict what is in front of them, and artistic or nefarious goals are served via such accuracy.

the authorities will need an accurate depiction of my face to hold me accountable.[14] And this is likewise the case for an audio recording of my boss making sexist or racist remarks.

Regarding descriptive accuracy, one might worry whether devices such as cameras really do get things descriptively right (even if we grant that there is less of a worry in this regard for audio-recording devices). For instance, Shen-yi Liao and Bryce Huebner (2021, 94) have argued that cameras have a "light-skin bias." The early development of Kodak film printing required that skin tones be matched to an image of a white model, with the result that "darker skin tones . . . [were] . . . over saturated, or under-lit, so the only images that . . . looked 'right' were images of light-skinned people" (95). Moreover, "light-skin bias . . . is not primarily a technical issue . . . . Film emulsions could have been designed that were more sensitive to a wider range of skin tones," (95) but they were not. Why? Because "deeply entrenched forms of racial ignorance and racial biases led the people who were developing emulsion technologies to ignore variations in skin tone, or assume that racialized differences were irrelevant to the design of this technology" (95; see also Roth 2009; Smith 2013; Benjamin 2019).

However, when we use cameras as a paradigm case of descriptive accuracy, we are considering an ideal—a system that *does* get things descriptively right (in a relevant regard). If a camera using Kodak film fails to accurately depict darker skin tones, then in that regard it will be descriptively *inaccurate*, notwithstanding its perhaps achieving descriptive accuracy in other regards.

Still, there are two important lessons. First, if a lack of descriptive accuracy in any particular respect is due to bias, as in the Kodak film case, we must be alive to the possibility that for language models, even when we *appear* to have descriptive accuracy about how people use language, we might not, at least in some relevant respect, perhaps due to bias in the training data.[15] Second, there are various respects in which a system or model might get matters *descriptively* right about how people use language, and different systems might have differing aims in this regard—for

[14] We will say more about why descriptively accurate information can be ethically useful in section 6. But note that even if knowing who perpetrated a wrongdoing *is* ethically useful information, we agree with Benjamin (2019) that often we have to ask prior questions about *whose* faces are being disproportionally recorded by cameras (e.g., members of overpoliced or oversurveilled communities), and in particular by devices such as CCTV. These considerations are consistent with our claims.

[15] This issue will be important later, when we discuss large language models. Some argue that models like BERT or GTP-2/3 not only reproduce bias (and are thereby normatively incorrect) but are descriptively inaccurate since the training data do not reflect marginalized voices (Bender, Gebru et al. 2021).

example, a model might aim for descriptive accuracy in one respect yet not in another. For example, word-embeddings models (which we will discuss in detail in section 4) aim at capturing how words cluster in a vector space based on how people actually use language, and these models might be assessed for descriptive accuracy in this respect. As we shall see, "woman" and "homemaker" do cluster together, whereas "woman" and "computer programmer" do not. Even so, such a model is clearly descriptively *inaccurate* in another sense, since these clusterings are descriptively inaccurate when it comes to these words' semantics—there is nothing about "woman," semantically, in virtue of which it clusters with "homemaker" (by contrast with "bachelor," which semantically clusters with the word "man").

One might think that whether a representation *is* descriptively accurate is a function of who we are and what we want to do with it—in which case, descriptive accuracy seems context-dependent. But this issue is more one of salience, not accuracy. Many representations *are* accurate even if some information they convey is not salient, and salience depends on the context. The family photograph I took in my front garden is descriptively accurate in depicting us smiling. In the context of sending a holiday card to friends, this information may be most salient. Yet the photograph might also accurately depict the hole in my eavestrough, and in a different context—such as getting a quote for fixing the hole—this descriptively accurate information will be of primary salience rather than the fact that we are smiling.[16]

Finally, note that descriptive accuracy does not require *perfect* descriptive accuracy. Many representations of aspects of the world are descriptively accurate even when they are less than perfectly accurate. That is the case even for scientific theories, including theories of physics (see, e.g., Hitchcock 2004, 10). Moreover, a visual representation can be accurate without being perfectly accurate. For instance, canine visual systems are optimized more for low-light conditions than color, by contrast with our own vision (Barber et al. 2020). Even so, dogs' vision is descriptively accurate in many respects and is often *more* descriptively accurate than ours. Finally, it is a truism that *all* models are inaccurate to some degree (see, e.g., Box 1976, 792) while still being usefully descriptively accurate (and thus also predictively or explanatorily useful).

Leaving aside descriptive accuracy, a system or its outputs are *normatively correct* if they reflect the world as it *ought* to be—in some respect and on some conception of how the world should be, even if that is not how it *is*. Language models are normatively correct if they reflect *language use* as it should be. For example, a model might be normatively correct in not giving sexist outputs. Such a system would not resolve the analogy task "Man is to computer programmer as woman is to *x*" by

[16] We thank Shelley Park (at the 2021 FSJAI Workshop) for prompting this clarification.

filling in "homemaker" for *x* (we will discuss this example in greater detail in section 4). By contrast, *normatively incorrect* models or outputs reflect language use as it ought not to be; such a system might fill in "homemaker" for *x* in our example.[17] For our purposes, morally relevant bias in a language model is sufficient for that model to be normatively incorrect, and in such cases these models reflect bias, as outlined in section 2, whereas normatively correct models or outputs do not (nor, consequently, do they perpetuate bias). A system might also be normatively correct in one regard (e.g., in not giving sexist outputs) yet incorrect in another (e.g., in giving racist outputs).[18]

Additionally, normative correctness can have various goals. Not reflecting ethically problematic biases might be one normatively correct goal, whereas other goals might include promoting outcomes related to social justice. We acknowledge these various goals as being normative in a relevant sense. Even so, our primary focus will be on the aim of not reflecting bias.

To see more clearly how descriptive and normative correctness relate to bias in NLP systems, let us consider the widespread NLP method of word-embeddings, which we will now describe.

4. Word-Embeddings: Descriptively Right, Normatively Wrong

Later, we will consider applications that rely on large language models. Yet, for now, in order to introduce the idea of bias in NLP models, we will outline the simpler case of word-embeddings.

Word-embeddings are learned representations of words that aim to approximate the semantic relationships between words through mathematical relationships between vectors (Mikolov et al. 2013). A word-embeddings model learns from vast language corpora produced by us, which are therefore examples of how we use language. By definition, a good set of representations will accurately reflect how language is, as a matter of fact, used by the sampled people. In this sense, word-embeddings do a remarkably good job of getting things descriptively right, since

[17] Normatively incorrect outputs *need* not be descriptively accurate, although often they are (see section 6.3).

[18] Arguably, the best way to ensure that language use is as it should be and that language models trained on such language use are normatively correct would be to have the nonlinguistic part of the world that the language use aims to describe—e.g., the proportion of women and men who are computer programmers—be as it should be. After all, people associate "he" with programmers more than "she" because traditionally most programmers have tended to be men. Yet over time, if more programmers were women, that would be less true. Normative correctness would be descriptively accurate.

they are usefully deployed for many purposes—for example, to successfully resolve analogy tasks. Yet there is a sense in which word-embeddings get things normatively wrong.

In their paper, "Man Is to Computer Programmer as Woman Is to Homemaker," researchers at Boston University and Microsoft Research (Bolukbasi et al. 2016) demonstrated that pretrained models such as word2vec seem sexist in certain ways. Just as the analogy relationships "Man is to woman as king is to queen" and "Sister is to woman as brother is to man" were captured by the models, so too were the sexist relationships "Man is to computer programmer as woman is to homemaker" and "Father is to doctor as mother is to nurse."[19]

The difficulty is that to work successfully, NLP systems relying on word-embeddings often *have* to learn the biases that exist in the corpora on which they are trained (Caliskan, Bryson, and Narayanan 2017, 186). These biases are expressed by us in the corpora, and to learn the relationships that actually exist between words in our uses of language, the models must learn biased relationships—including sexist relationships. Biases in the corpora on which the models are trained will thus naturally be captured in the geometry of the word-embeddings vector space. Worse, according to Bolukbasi and colleagues' assessment, "The blind application of machine learning runs the risk of *amplifying* [emphasis added] biases present in data" (Bolukbasi et al. 2016).[20]

Still worse, word-embeddings models often get things normatively wrong *precisely because*—and to the extent to which—they get things descriptively right about people's language use. That is, the success of systems using word-embeddings can *require* reproducing bias, as when they solve analogy tasks like "Man is to

[19] For criticism of Bolukbasi et al. (2016), see Schluter (2018), Nissim, van Noord, and van der Goot (2019), and footnote 24 below.

[20] Bias is amplified in two ways. First, sexist bias (for example) in a language model is acquired from training data that include expressions of sexism from (*p* number of) individuals expressing the bias. This model may then be used for various purposes in the world such that its outputs reach many more people (*p + n*) than those people (*p*) who expressed the biases that influenced the training data. Mere repetition is widely known to increase people's tendency to believe claims or to reproduce bias, including in written language (e.g., Lacassagne, Béna, and Corneille 2022). In this way, bias gets amplified in the population due to biased models. Second, this larger pool of individuals (*p + n*) is therefore more likely to reproduce examples of language containing bias than it would have been otherwise, and these examples may be used as input for training further models. Thus, the bias gets amplified to a wider range of models than it would have been otherwise; see Bolukbasi et al. (2016) for a similar example.

computer programmer as woman is to *x*" by filling in "homemaker" for *x.* A model gets things normatively wrong in producing this output, despite its being descriptively right in reflecting actual language use. And it gets things normatively wrong *because* and *to the extent to which* it gets things descriptively right in this way.[21]

The ideal solution might be to debias ourselves. In that case, we would produce language corpora that do not contain biases, the NLP models would be trained on such data, and all would be well. Yet that is not a practical suggestion, much as it might be nice to imagine living in such a world. Or we might debias the models that have acquired biases. As mentioned earlier, it can be difficult, as cognitive science shows, to eliminate biases from human cognition. We think that it is no easier to debias NLP models, which raises a practical problem, which we outline in section 5, and an ethical dilemma, which we outline in section 6.

5. Debiasing: Practical Problems

Word-embeddings and large language models are trained using neural networks whose objective function is to minimize error on a particular task, often the ability to predict each word in a text given the surrounding words (see, e.g., Mikolov et al. 2013). That is the training task. When we consider downstream tasks, we are taking a model that has already been trained on a training task using a massive language corpus and fine-tuning it or using it *as is* on a new task. The new task could be text classification, which might then be used in a further application—for example, a content-management system that adds keywords and categories to content.

This case is an example of transfer learning, where learning from one task on a massive corpus is transferred to another task on a different, smaller corpus. Often, the so-called pretrained model is made publicly available by those who trained it and can end up being used in countless downstream applications that the original developers of the model might never know about. That is what happened with word2vec, the word-embeddings model we mentioned above (Mikolov et al. 2013), and BERT, one of the first large language models to be made publicly available (Devlin et al. 2019). The focus on debiasing has tended to be on this initial stage of the training pipeline.[22] The resulting models are subsequently released for use in various applications; if we debias these pretrained models, then the applications built

[21] We acknowledged in section 3 that such a model is descriptively *inaccurate* regarding these words' semantics—there is nothing about the word "woman," semantically, in virtue of which it clusters with "homemaker."

[22] Here, word-embeddings models *are* the initial stage of the training pipeline and can subsequently be used in downstream applications (Mikolov et al. 2013). Blodgett et al. (2020) surveyed 146 papers on bias in NLP systems and 54 of these papers specifically dealt with bias in word-embeddings systems.

downstream of them will also be debiased unless bias is introduced somehow later. Consequently, debiasing the models is seen as an attractive solution to many researchers. However, we think that researchers would do better to focus on debiasing the outputs of models rather the models themselves. In turn, we claim that given the serious practical difficulties with debiasing models, it behooves us to consider whether the biased models we are left with are always bad. We will argue that they are not (in, e.g., section 7 below).

In their previously mentioned paper, Bolukbasi and colleagues (Bolukbasi et al. 2016) introduce a method they call "hard-debiasing," which aims to reduce sexist bias within an embeddings vector space without compromising the overall structure of that space or, therefore, its usefulness. They claim that their method can "significantly reduce gender bias in embeddings while preserving [their] useful properties such as the ability to cluster related concepts and to solve analogy tasks" (Bolukbasi et al. 2016, 1).[23, 24]

We doubt whether word-embeddings can be completely debiased using this method. What Bolukbasi and colleagues call hard-debiasing works largely by maintaining humans in the loop at multiple steps to identify and compensate for biases picked up by the model. This feature is partly attractive, since humans in the loop are frequently practically and ethically beneficial (especially when the stakes are high, as in medical applications, since the instance of harmful errors will thereby be lowered; arguably, humans must always be kept in the loop in such cases). Yet as applied to word-embeddings, the end product is less a well-oiled machine that works by itself and more an old TV that must be continually retuned for clear reception.[25] The retention of humans in the loop as a means of policing a system's acquisition of bias also raises worries about *implicit* bias in human cognition. In particular, people

[23] There are also technical aspects to Bolukbasi and colleagues' hard-debiasing method that we do not discuss.

[24] A word about analogy tasks: Nissim, van Noord, and van der Goot (2019) outline various problems with using analogy tasks for bias detection. At their worst, some algorithms ignore input vectors, so that "A is to B as C is to *x*" is often prevented from returning A, B, or C for *x* (Nissim, van Noord, and van der Goot 2019, 491–92; cf. Schluter 2018). Nissim and colleagues think using analogy tasks to detect bias in embeddings results in claims about biases in models where the evidence does not warrant it and may even conceal existing biases.

[25] A solution might be not to use such models. We are concerned with what to do given that they *are* used.

may unconsciously *add* biases even if they succeed in eliminating others (see Caliskan, Bryson, and Narayanan 2017).[26]

Finally, it is unclear whether we can realistically remove *all* morally relevant biases from a word-embeddings model while preserving its usefulness. Any tweaks made to a pretrained word-embeddings model to increase its *normative* correctness in a particular regard will result in a less *descriptively* accurate model regarding actual language use, at least in cases where the models reflect bias. Many applications will be negatively affected by this reduced accuracy. For example, when pretrained models are used in sentiment analysis and text classification, which *rely* on descriptively accurate representations, these models seem a poor fit for those tasks.

Such practical problems with debiasing suggest that a more effective way to control for the negative effects of bias in word-embeddings might instead be to see how bias manifests itself in a system's outputs and control for bias there to reduce the negative effects.

6. Applications

Controlling for bias at the output stage means we can retain descriptive accuracy. Exactly how we decide to control for bias in this way will, of course, depend on the application in question. Let us therefore consider some applications.

6.1. Translation

Machine translation is the oldest NLP task and is often regarded as one of the field's major successes. Deep-learning techniques similar to those employed in the training of large language models such as word2vec and BERT are employed in today's most commonly used machine-translation systems, such as Google Translate. One difference is that these systems are trained end-to-end, meaning that there is no breaking up of the problem into separate steps whereby a language model is first trained on a separate training task and only later fine-tuned for the task of translation. Instead, the translation task *is* the training task (Wu et al. 2016).

In spite of their successes, machine-translation systems still encounter trouble with pronouns. Douglas Hofstadter (2018) has written about the "shallowness" of machine translation, pointing to such systems' inability to *understand* what they are translating. Many of Hofstadter's examples are of mistranslating pronouns. The reason pronouns are so difficult is that to translate a pronoun correctly, a system must

[26] Bolukbasi et al. (2016) concede that their method is limited. Gonen and Goldberg (2019) argue that even the technical aspects of Bolukbasi and colleagues' method serve primarily to mask the problem of gender bias without actually eliminating it. They conclude that, "While the bias is indeed substantially reduced …, the … effect is mostly hiding the bias, not removing it" (Gonen and Goldberg 2019, 609).

know what it refers to. With gendered pronouns, the problem is clear. Some languages do not have gendered pronouns, and those that do can behave very differently—for example, with possessive pronouns in French, the gender agrees with the noun, not the possessor (as in English). Hofstadter illustrates the problem as follows:

> I began my explorations . . . using the following short remark . . .
>
> > In their house, everything comes in pairs. There's **his** car and **her** car, **his** towels and **her** towels, and **his** library and **hers**.
>
> . . . Here's what Google Translate gave me [translating into French]:
>
> > Dans leur maison, tout vient en paires. Il y a **sa** voiture et **sa** voiture, **ses** serviettes et **ses** serviettes, **sa** bibliothèque et **les siennes**.
>
> The program fell into my trap, not realizing, as any human reader would, that I was describing a couple, stressing that for each item *he* had, *she* had a similar one. . . . The deep-learning engine used the word *sa* for both "his car" and "her car," so you can't tell anything about either car owner's gender. Likewise, it used the genderless plural *ses* both for "his towels" and "her towels," and in the last case of the two libraries, his and hers, it got thrown by the final *s* in "hers" and somehow decided that that *s* represented a plural ("*les siennes*"). Google Translate's French sentence missed the whole point. (Hofstadter 2018; bold emphasis added)

Data-driven approaches seem unlikely to solve this sort of problem. Yet gendered pronouns are an obvious place where bias reveals itself in machine translation. For example, Turkish does not have gendered pronouns yet there are numerous examples of Turkish sentences translated into English by Google where a genderless pronoun *becomes* gendered according to whether the noun is a word for a stereotypically male or instead female profession: "he" for doctor, "she" for nurse, and so forth. Thus, "O bir doktor" will be translated as "He is a doctor," even though the Turkish pronoun "O" is ungendered and can be translated as either "he" or "she." In 2018, Google introduced gender-specific translations in order to avoid gender bias in Google Translate. As a result, when you enter "O bir doktor," the system now offers a

choice—either "he" or "she"—and it lets the end user decide on the correct pronoun.[27]

But when context is provided, Google Translate gets it wrong. If you are simply translating "O bir doktor," there is no wrong answer, since there is no context provided; either "he" or "she" is acceptable. However, with a real piece of text, Google's system reveals its bias. For example, it will provide the following (correct) translation from English into Turkish:

> Did you meet John's brother? **He** is a nurse.
> John'un erkek kardeşiyle tanıştın mı? **O** bir hemşire.

Yet when translating this Turkish sentence back into English, the output is:

> Did you meet John's brother? **She** is a nurse.

This result is a double fail: The translation is incorrect (even if this usage of "she" in relation to "nurse" remains descriptively accurate about people's actual language use), and the output is biased. Moreover, it reveals the changes that Google implemented as relatively superficial and insubstantial.

6.2. Assorted NLG Tasks

Translation is an example of a *natural-language-generation* (NLG) task, where the output is a sentence or sentences in a natural language. Other examples of NLG tasks include speech recognition, image captioning, and dialogue systems. One naïve suggestion for systems that perform NLG tasks might be that they should *never* give biased outputs, such as sexist or racist outputs.

When we look at particular applications, we can see why this suggestion is naïve. If a speech-recognition system receives as audio input the sentence "A proper wife should be as obedient as a slave," the only appropriate output is text of these words.[28] There is no scope for debiasing without undermining the task that the system is designed to perform. Descriptive accuracy trumps normative correctness in this case, since, as with a camera, the functions of a speech-recognition system require descriptive accuracy.

---

[27] We made this suggestion in a talk we gave at Google in 2018. Much as we would like to take credit for the change, it was no doubt something that the people at Google had already been discussing long beforehand.

[28] This phrase, "A proper wife should be as obedient as a slave," often appears online as attributed to Aristotle. It seems to be a broad paraphrase of scattered parts of Aristotle's *Oeconomica*, Book III, section 1, paragraph 2.

In translation tasks, there may not be only one acceptable output, since for any phrase we input, there may be multiple acceptable ways of translating it into another language. Even so, there appears to be limited scope for debiasing or imposing normative correctness. The output must be an accurate translation of the sentence—for example, "A proper wife should be as obedient as a slave"—into the target language, however sexist or morally objectionable that sentence is. Again, we cannot impose normative correctness without undermining accuracy.

With image captioning, we begin to see scope for imposing normative correctness. Clearly, there are numerous suitable captions that a system might suggest for any image. For a photo of a woman kitesurfing, a sexist person might suggest the caption, "A disobedient wife." Yet we would hope that no image-captioning system would ever do so. Thus, there seems to be a normative constraint on the output. Obviously, we do not want an image-captioning system to label a photo of a kitesurfing woman with "Solar system," or "Proton," since these captions are *descriptively* inaccurate. Assuming the aim is descriptive accuracy regarding the image's content, the caption must accurately describe what is in the image. Because images are typically more informationally rich than linguistic descriptions of them, most captions (even if they are *partly* descriptively accurate) will only be *minimally* accurate. As a result, "Woman sailing" might be minimally accurate, while "Woman kitesurfing" might be better. But "Woman kitesurfing on a Slingshot kiteboard in light seas and moderate winds" might be *too* accurate for most contexts. In any case, "Disobedient wife" is *not* an accurate description in virtue *merely* of the informational content of the image. This caption might be descriptively accurate of the *situation* being depicted, were we to learn (for example) that *this* kitesurfing woman in fact disobeyed her sexist husband's command never to kitesurf, since the husband believes wives—being women—should not kitesurf. Once we understand the caption "Disobedient wife" in a pejorative sense—as implying something like, "This woman, in virtue of kitesurfing, is behaving as no woman should behave"—we can see it would be normatively incorrect to allow the caption. Most image-captioning systems would be unlikely to provide this caption in any case, since they would not have encountered it in their training data. Beyond these considerations, we also want a system to be normatively correct in not labeling the image with "A disobedient wife" even *if* the system *had* been exposed to this label in its training, since the label is sexist—that is, *normatively* incorrect.

Finally, consider dialogue systems. Here, there can be appropriate and inappropriate ways to respond to a question, even descriptively. In answer to the question, "What is the weather like in Sydney?" a system can reply "Sunny" or "Beautiful," but presumably not "Subatomic" or "Bread." Yet if someone asked, "How obedient should a proper wife be?" the system should *never* answer, "A proper wife

should be as obedient as a slave."[29] What is an acceptable output in the case of speech recognition should never be output by a dialogue system. Here, we reach maximal scope for imposing normative correctness on a system's outputs.

The upshot is, first, that it can be preferable—not only practically but ethically (we discuss why this is an ethical issue in sections 7 and 8)—to control for biases in ML models and systems at the output stage rather than to try to debias a model and expect to deploy it for downstream tasks in ways that will not reproduce bias. Second, even when we focus on controlling for biases at the output stage, NLG applications must be treated differently, including from an ethical perspective. In some cases we can reduce bias; in others we cannot.

6.3. Search

Search is a different application. In considering sources of bias in ML systems, Harini Suresh and John Guttag (2020) discuss two that are relevant to both search and our purposes in this paper—that is, *historical* and *representational* bias (see also Fazelpour and Danks 2021). The latter bias results from imperfect measuring or sampling. By contrast, according to Suresh and Guttag (2020, 4), "historical bias arises even if the data is perfectly measured and sampled, if the world *as it is* or *was* leads a model to produce outcomes that are not wanted." They provide an example of this bias:

> In 2018, 5% of Fortune 500 CEOs were women . . . . Should image search results for "CEO" reflect that number? Ultimately, a variety of stakeholders, including affected members of society, should evaluate the particular harms that this result could cause and make a judgment. This decision may be at odds with the available data even if that data is a perfect reflection of the world. Indeed, Google has recently changed their Image Search results for "CEO" to display a higher proportion of women." (Suresh and Guttag 2020, 5)

In our terms, Suresh and Guttag claim that descriptive accuracy ("the world *as it is* or *was*") conflicts here with normative correctness (i.e., we get "outcomes that are not wanted"). That is to say, imposing normative correctness by displaying a higher proportion of women in the image search results may be descriptively at odds with

[29] If we asked, "How obedient *does Aristotle say* a proper wife should be?" the appropriate answer might be "*Aristotle says* that a proper wife should be as obedient as a slave." Yet when asking the system itself, it seems clear that imposing normative correctness by forbidding such a response is justified.

the data even if the data are "a perfect reflection of the world."[30] By altering the results, Google would seem to sacrifice descriptive accuracy in favor of normative correctness, by glossing over the actual disparity between women and men who are CEOs. This inaccuracy highlights an ethical problem. To possess such descriptively accurate information is often to acquire *ethically useful information*—in this case, the information might be used as evidence for the claim that there is indeed a disparity between women and men CEOs, perhaps in the course of redressing the disparity. However, once Google alters the results as described, one effect is that we lack access to this descriptively accurate and ethically useful information and so injustices of this sort might go underrecognized.

Moreover, it is implicit in Suresh and Guttag's discussion of their example that there is a value both to descriptive accuracy and to normative correctness, although Suresh and Guttag are unclear about whether they think it can be more valuable to prioritize one form of accuracy or correctness over the other. After all, while *some* ethical purposes might best be served by Google's decision to change its results for "CEO" to display more images of women (thereby making the results more normatively correct), *other* ethical purposes may be directly undermined and would be better served by descriptively accurate results; so, Google's strategy may be ethically undesirable in other respects. That is not to say Google made the wrong decision. It is only to say that there might be a cost to prioritizing *either* descriptive accuracy *or* normative correctness in search, as with NLP systems relying on word-embeddings.

In fact, we have reservations about whether our ethical dilemma applies in Suresh and Guttag's CEO case, or in search more widely. That is because it is unlikely that Google Search (or search engines generally) actually provide us with results that are descriptively accurate in the first place. For example, Safiya Umoja Noble (2018) outlines how Google's search results can be racist in ways that are driven by internal factors, such as an agenda built into Google's algorithms, or by external factors, such as how savvy users can deliberately influence results.

To the extent that search results are *not* descriptively accurate, our ethical dilemma does not arise (as we explain in detail in section 7). We confront the dilemma only when we have descriptively accurate outputs that are normatively incorrect, and the incorrectness is due to the accuracy. In such cases, the dilemma consists in our having to decide whether to prioritize descriptive accuracy (including its ethical

[30] In another widely discussed case, *The Guardian* reported in 2016 how Twitter user Kabir Alli conducted a Google image search using the phrases "three black teenagers" and "three white teenagers" (Allen 2016). Results were very different. The search for "three black teenagers" returned mostly police mugshots while for "three white teenagers" it gave stock photographs of smiling white youths.

usefulness) at the expense of normative correctness or vice versa. Even if search results are *normatively incorrect* (as in Suresh and Guttag's case) unless they are also *descriptively accurate,* there is no dilemma (as we explain in sections 7 and 8 below); we must simply address the matter by trying to attain normative correctness.[31]

Even so, in an idealized case where search results did reflect descriptive accuracy and were normatively incorrect, our dilemma would frequently arise. After all, descriptively accurate information can—as in the case of word-embeddings models, for instance—be ethically useful in studying biases (as expressed in language), how they are perpetuated, and how to address them.

7. The Ethical Dilemma

It can sometimes be ethically valuable *not* to debias NLP models or their outputs. For one thing, not debiasing enables us to uncover biases and can thereby prevent us from thinking that all is right with the world (in some regard) when it is not. However, a lot depends on the application under consideration. With speech recognition, there is no room for imposing normative correctness, and doing so might even be ethically problematic. Imagine we are listening to Donald Trump making a speech on TV and our speaker is not working. Instead, we are relying on a live-captioning feed. We would be poorly served by this system were it to impose normative correctness by producing an eloquent discourse on feminism. We need to know what Trump is actually saying—unpalatable as it might be—in order to hold him accountable.

At the other end of the spectrum is search. If search results do *not* reflect descriptive accuracy (see, e.g., Noble 2018), the ethical dilemma that we outline does not arise. If search results *were* somehow to reflect descriptive accuracy but were normatively incorrect, we might confront the dilemma. We would then be faced with two options: either prioritize descriptive accuracy over normative correctness, which has the potential cost of perpetuating or amplifying bias, or instead prioritize normative correctness, with the potential cost of withholding ethically useful information even if we thereby avoid perpetuating or amplifying bias.

Recall that *descriptive accuracy*, as we use this term, means *sufficient* descriptive accuracy in *a relevant regard* (all models are inaccurate in some—and presumably many—regards, even with respect to a target phenomenon). Without descriptive accuracy we cannot proceed usefully and there is no dilemma. Call a situation *Dilemma-land* when we have a descriptively accurate model that, in virtue of its descriptive accuracy, gives normatively incorrect (i.e., biased) outputs. By contrast, call a situation *Utopia-land* when we have a descriptively accurate model

[31] Alternatively, we could strive for descriptive accuracy, in which case our dilemma might apply after all.

that gives normatively correct outputs. There is no way of getting to Utopia-land until people no longer express biases (e.g., sexism or racism) in the language corpora on which we train models.

What if we have descriptively *inaccurate* models? Perhaps, in that case, we only need *normative* correctness. However, an inaccurate model may fail to sufficiently perform the task it is meant to perform, since inaccuracy often undermines usefulness. Achieving normative correctness without descriptive accuracy would, in any case, put us in *Fantasy-land*; if a model exhibits normative correctness but is descriptively inaccurate, it is out of touch with reality and is thus a fantasy (although sometimes, as we acknowledge below, we may have reason to prefer such fantasies to descriptive accuracy—for example, normatively correct outputs may serve to work against amplification of bias in some cases, even if the model is descriptively inaccurate).

Achieving normative incorrectness *and* descriptive inaccuracy puts us instead in *Disaster-land*. Although it may be possible to get from Fantasy-land or Disaster-land to Utopia-land, it is highly unlikely. (We leave confirmation to our readers.) More likely is that we simply cycle back and forth between Fantasy-land and Disaster-land by achieving normative correctness/incorrectness in our model yet without descriptive accuracy. Either way, until we have descriptive accuracy, we cannot get to either Dilemma-land or Utopia-land; and (as indicated above) Dilemma-land remains a more likely destination than Utopia-land anyhow.

Only in Dilemma-land does the dilemma arise, since for it to arise a model must exhibit descriptive accuracy and, in virtue of this descriptive accuracy, produce normatively incorrect outputs (if it gives normatively incorrect outputs resulting from something *else*, that is a different matter that must be addressed on its own terms). In Dilemma-land, we must decide whether to prioritize descriptive accuracy or instead aim for normative correctness by debiasing. Debiasing may undermine the usefulness of the model and might undermine descriptive accuracy and thus withhold ethically useful information. Yet prioritizing descriptive accuracy runs the risk of perpetuating or amplifying the relevant biases. If a model exhibits descriptive inaccuracy, there is simply no dilemma to confront given that we would never want to prioritize descriptive inaccuracy over either normative correctness or incorrectness.

Our aim is to draw attention to this dilemma. Language models and various applications (even potentially including search) that rely on them can be used for different purposes. Sometimes these purposes are best served by having the model reflect descriptive accuracy even when that entails reflecting bias. Those purposes will, moreover, often be served precisely *because* the biased results provide us with ethically useful information. However, we do run the risk of amplifying bias. In other cases, an application might be used for a different purpose, such that it is more

important to avoid amplifying bias than to retain ethically useful information. Then, we may prefer normatively, not descriptively, correct results.

Either way, there is a cost. We also maintain that there is sometimes an ethical value to the less intuitive horn of the dilemma, on which we retain ethically useful information by not imposing normative correctness. Even so, we recognize that in doing so we can amplify bias.

8. Rethinking the Debate about Bias in NLP

Our way of distinguishing normative correctness and descriptive accuracy adds nuance to recent debates about algorithmic bias in NLP and can help us to understand these debates better.

For instance, as outlined above, Suresh and Guttag's (2020) historical bias can be understood as arising when we have descriptively accurate outputs (regarding people's use of language) about how the world is yet these outputs are normatively incorrect. Thus, our distinction can help to clarify the points of contention among these debates' participants and the implications of prioritizing one form of accuracy over another. Once we acknowledge the ethical dilemma that arises when descriptive and normative accuracy conflict in the way we have outlined, we can see clearly the ethical costs of prioritizing one form of accuracy over the other.

To take another example, Emily Bender, Timnit Gebru, and colleagues argue that large language models like BERT or GPT-2/3 not only reproduce bias but do *not* reflect "the world *as it is*" (Suresh and Guttag 2020, 4), since the corpora on which they are trained do not reflect marginalized voices (Bender, Gebru et al. 2021).[32] Thus, one cannot argue on the basis of a model's descriptive accuracy for not debiasing it if it is not descriptively accurate.

Part of the problem is that while "user-generated content sites like Reddit, Twitter, and Wikipedia present themselves as open and accessible to anyone, there are structural factors including moderation practices which make them less welcoming to marginalized populations" (Bender, Gebru et al. 2021, 613), resulting in marginalized voices' being underrepresented in the models. Simultaneously, "white supremacist and misogynistic, ageist" voices tend to be "overrepresented in the training data" (613) due to how the data are filtered.

Bender/Gebru and colleagues' description of this problem can be usefully framed in terms of normative and descriptive correctness. Large language models, Bender/Gebru and colleagues appear to say, are not only normatively incorrect but also descriptively inaccurate in important ways, and they are normatively incorrect partly *because* they are descriptively inaccurate.

[32] Bender and Gebru are listed as joint lead authors on their 2021 paper. We have slightly modified citation format for that paper to reflect this.

Even criticisms of Bender/Gebru and colleagues' claims can be usefully framed in terms of normative correctness and descriptive accuracy. For instance, Yoav Goldberg (2021) claims that Bender/Gebru and colleagues "suggest that good (= not dangerous) language models are . . . models which reflect the world as they think the world should be," yet "an alternative view by which language models should reflect language as it is being used in a training corpus is at least as valid, and should be acknowledged." Goldberg picks up here on an ambiguity in how "good" (or its cognates) can be used in this context—that is, as picking out normative correctness ("not dangerous") or descriptive accuracy ("language as it is being used"). In our terms, Goldberg claims that Bender/Gebru and colleagues fail adequately to acknowledge their preference for prioritizing the normative over the descriptive, and in a particular way. Indeed, Goldberg thinks that Bender/Gebru and colleagues prioritize normative correctness in a way that can result in our adopting certain views on ethical issues even when there is legitimate debate about whether they are the right views to adopt. As Goldberg (2021) puts it, the fact that there is ongoing debate on such issues "should be made explicit" by Bender/Gebru and colleagues. Goldberg also thinks there can be practical and ethical reasons for prioritizing descriptive accuracy over normative correctness, contra what he takes Bender/Gebru and colleagues to implicitly maintain.

A central target of Bender/Gebru and colleagues' concern seems not to be what Suresh and Guttag call historical bias but rather what they call *representation bias*, which occurs "when certain parts of the input space are underrepresented" (Suresh and Guttag 2020, 5). For example, with language models, "datasets collected through smartphone apps can under-represent lower-income or older groups, who are less likely to own smartphones" (Suresh and Guttag 2020, 5). This form of bias is one of the central focuses of Bender/Gebru and colleagues' attention, rather than historical bias, given that a central claim of theirs is that often the relevant data have *not* been perfectly measured and sampled. By contrast, historical bias arises even when the data *have* been "perfectly measured and sampled" and "the world *as it is* or *was* leads a model to produce outcomes that are not wanted" (Suresh and Guttag 2020, 4).

Our normative/descriptive distinction again helps to clarify the points of disagreement and agreement between Bender/Gebru and colleagues and their critics such as Goldberg, while additionally clarifying how language models can involve representation bias in Suresh and Guttag's sense. Bender/Gebru and colleagues' claim is that large language models are descriptively inaccurate relative to the wider population of language users, as a result of their being built around an imperfect measuring and sampling of data. Such models can be normatively incorrect in ways

that *result from* such inaccuracy.[33] Thus, Suresh and Guttag's representation bias is one way in which a model can be *descriptively* inaccurate. When such inaccuracy is distinguished from *normative* incorrectness, the points of disagreement among Goldberg and Bender/Gebru and colleagues become clearer.

It might be suggested that Bender/Gebru and colleagues underestimate the extent to which language models like GTP-2/3 *are* descriptively accurate. After all, even if they are trained on datasets that come in significant part from platforms from which marginalized voices have been excluded, and so the models are descriptively inaccurate at least to that extent, the models might still be descriptively accurate about how language is used *on those platforms*.

Bender/Gebru and colleagues' response might be that the aim of descriptive accuracy is to capture patterns that exist in the entire population (or most of it), not simply a subpopulation. For large language models, the aim (in that case) would be that models accurately reflect how humans in general, including those in marginalized groups, use language. If we take language from only a specific subpopulation, we cannot make claims about descriptive accuracy (except, perhaps, regarding that specific subpopulation's use of language). The most obvious bias here would seem to be representational bias, given that the model is only representative of a specific subgroup while ignoring others. Even practically, the model's usefulness for downstream tasks on texts written by other language users will thus be limited.

Yet suppose we trained a hypothetical model on *all* of the language ever produced by people across *all* groups.[34] We might then have descriptive accuracy. Even so, bias might remain—not representational bias but historical bias. In our terms, we would have descriptive accuracy yet normative incorrectness (and the normative incorrectness would be *due* to the descriptive accuracy). For example, we would still have the problem of gendered pronouns in translation. If people speak as though there are more female than male nurses, the pronoun "she" will still show up more frequently near "nurse" than "he" in the relevant vector space, even across the enormous dataset of all the language ever produced by all groups. Likewise, if people speak as though there are more male than female programmers, "he" will show up more frequently near "computer programmer" than "she." This bias comes from descriptive, historical facts about language use, not from how the dataset was built.

---

[33] Bender, Gebru et al. (2021) thereby seem to be saying that *normative incorrectness* can result from *descriptive inaccuracy*. We do not deny this, although our focus is instead on how descriptive accuracy can conflict with normative accuracy—i.e., how *normative incorrectness* can result from *descriptive accuracy*.

[34] Falbo and LaCroix (2022) note that minority groups often "code-switch" by conforming to majority norms of expression; thus, the content they *do* generate might *still* not reflect marginalized voices.

The distinction between representation bias and historical bias is also important when we consider whether there can be a value to *not* debiasing models. Historical bias often indicates something interesting about the world, whereas representational bias might—in the worst cases—only indicate lazy data-gathering. The more descriptively accurate a model is, the more useful it will be for downstream tasks and the more accurately it will reflect historical (and current) biases in language use. The choice to debias must *start* with a descriptively accurate model. In that case, we encounter the dilemma: Either we debias the accurate model and avoid amplifying bias yet lose accuracy and ethically useful information, or we do not debias and we use the model to learn about biases, for example, but risk amplifying bias.

9. Conclusion

What should we do about bias in NLP systems? It depends, of course, on the *source* of such bias. But even for bias from properly sampled data, it also depends on what we *want* from such systems. Do we want them to be more like cameras or more like people? We expect cameras to get things descriptively right in certain relevant regards, but not normatively right. Yet we expect people to get things normatively right, even if they get things descriptively wrong. For example, we do not excuse someone for making a sexist comment even if their making that comment is a descriptively accurate reflection of how language is actually used in their language community—we want the person to get things normatively right, *not* descriptively right.

Should we expect NLP systems to get things normatively right, even if they get things descriptively wrong? Such systems could work like the iBlur camera that we described in section 3, or like the speech-filter in the TV show *The Good Place*, where Eleanor says things like "Holy motherforking shirtballs!" despite presumably trying to say something else. This filter "translates out" things that we do not want to hear or are not permitted to hear. Alternatively, is it enough for NLP systems to be like cameras in getting things descriptively right, so that they reflect the actual world and let us decide what to do about it in particular contexts?

We maintain that there can be a value to having NLP systems function more like cameras. For one thing, it can provide ethically useful information, which we may even have an ethical obligation to provide. Then we face a dilemma, since allowing a system to get things descriptively right by not debiasing it can perpetuate or amplify bias. We think that deciding which way to go can only be done on a case-by-case basis, by weighing the ethical costs and benefits.

Our primary aim has been to draw attention to the dilemma, which arises once we acknowledge that there can often be a value, however counterintuitive, to not debiasing. We have also argued that it is frequently best to control for the negative effects of bias at the output stage rather than by debiasing models themselves. Yet

even then, we cannot easily avoid the dilemma, since we must still decide whether normative correctness should overrule descriptive accuracy in cases where descriptive accuracy provides us with ethically useful information.

Finally, our distinction between normative correctness and descriptive accuracy provides help in better understanding recent debates about algorithmic bias in natural-language processing. In particular, the question of whether to debias helps us to better understand these debates by bringing the consequences of adopting one view over another into clearer focus.

References

Allen, Antoine. 2016. "The 'Three Black Teenagers' Search Shows It Is Society, Not Google, That Is Racist." *Guardian*, June 10, 2016, https://www.theguardian.com/commentisfree/2016/jun/10/three-black-teenagers-google-racist-tweet.

Barber, Anjuli L. A., Daniel Mills S., Fernando Montealegre-Z, Victoria F. Ratcliffe, Kun Guo, and Anna Wilkinson. 2020. "Functional Performance of the Visual System in Dogs and Humans: A Comparative Perspective." *Comparative Cognition & Behavior Reviews* 15:1–44. https://doi.org/10.3819/CCBR.2020.150002.

Basu, Rima. 2020. "The Specter of Normative Conflict? Does Fairness Require Inaccuracy." In *An Introduction to Implicit Bias: Knowledge, Justice, and the Social Mind*, edited by Erin Beeghly and Alex Madva, 191–210. New York: Routledge.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *FaccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. New York: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922.

Benjamin, Ruha. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. Medford, MA: Polity Press.

Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. "Language (Technology) Is Power: A Critical Survey of 'Bias' in NLP." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, 5454–76. Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485.

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings." arXiv preprint, arXiv:1607.06520 [cs.CL]. https://doi.org/10.48550/arXiv.1607.06520.

Box, George E. P. 1976. "Science and Statistics." *Journal of the American Statistical Association* 71 (356): 791–99.

Brownstein, Michael. 2018. *The Implicit Mind: Cognitive Architecture, the Self, and Ethics*. New York: Oxford University Press.

Brownstein, Michael, and Jennifer Saul, eds. 2016a. *Implicit Bias and Philosophy: Volume I, Metaphysics and Epistemology*. Oxford: Oxford University Press.

———, eds. 2016b. *Implicit Bias and Philosophy: Volume 2, Moral Responsibility, Structural Injustice, and Ethics*. Oxford: Oxford University Press.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. "Semantics Derived Automatically from Language Corpora Contain Human-Like Biases." *Science* 356, no. 6334 (April 14): 183–86. https://doi.org/10.1126/science.aal4230.

Deery, Oisín. 2021. *Naturally Free Action.* Oxford: Oxford University Press.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, edited by Jill Burstein, Christy Doran, and Thamar Solorio, 4171–86. Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423.

Falbo, Arianna, and Travis LaCroix. 2022. "Est-ce que Vous Compute? Code-Switching, Cultural Identity, and AI." *Feminist Philosophy Quarterly* 8 (3/4): article 9.

Fazelpour, Sina, and David Danks. 2021. "Algorithmic Bias: Senses, Sources, Solutions." *Philosophy Compass* 16, no. 8 (August): e12760. https://doi.org/10.1111/phc3.12760.

Gendler, Tamar Szabó. 2011. "On the Epistemic Costs of Implicit Bias." *Philosophical Studies* 156, no. 1 (October): 33–63.

Goldberg, Yoav. 2021. "A Criticism of 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?'" GitHub post, January 23, 2021. https://gist.github.com/yoavg/9fc9be2f98b47c189a513573d902fb27.

Goldin, Claudia, and Cecelia Rouse. 2000. "Orchestrating Impartiality: The Impact of 'Blind' Auditions on Female Musicians." *American Economic Review* 90, no. 4 (September): 715–41. https://doi.org/10.1257/aer.90.4.715.

Gonen, Hila, and Yoav Goldberg. 2019. "Lipstick on a Pig: Debiasing Methods Cover Up Systematic Gender Biases in Word Embeddings but Do Not Remove Them." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, edited by Jill Burstein, Christy Doran, and Thamar Solorio, 609–14. Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1061.

Hitchcock, Christopher. 2004. "Introduction: What Is the Philosophy of Science?" In *Contemporary Debates in the Philosophy of Science*, edited by Christopher Hitchcock, 1–19. Malden, MA: Blackwell.

Hofstadter, Douglas. 2018. "The Shallowness of Google Translate." *Atlantic*, January 30, 2018. https://www.theatlantic.com/technology/archive/2018/01/the-shallowness-of-google-translate/551570/.

Hu, Lily, and Yiling Chen. 2017. "Fairness at Equilibrium in the Labor Market." arXiv preprint, arXiv:1707.01590 [cs.GT]. https://doi.org/10.48550/arXiv.1707.01590.

Kehl, Danielle, Priscilla Guo, and Samuel Kessler. 2017. "Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing." Responsive Communities Initiative/ Berkman Klein Center for Internet & Society, Harvard Law School. http://nrs.harvard.edu/urn-3:HUL.InstRepos:33746041.

Lacassagne, Doris, Jérémy Béna, and Olivier Corneille. 2022. "Is Earth a Perfect Square? Repetition Increases the Perceived Truth of Highly Implausible Statements." *Cognition* 223 (June): 105052. https://doi.org/10.1016/j.cognition.2022.105052.

Levy, Neil. 2017. "Implicit Bias and Moral Responsibility: Probing the Data." *Philosophy and Phenomenological Research* 94, no. 1 (January): 3–26. https://doi.org/10.1111/phpr.12352.

Liao, Shen-yi, and Bryce Huebner. "Oppressive Things." *Philosophy and Phenomenological Research* 103, no. 1 (July): 92–113. https://doi.org/10.1111/phpr.12701.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." arXiv preprint, arXiv: 1301.3781 [cs.CL]. https://doi.org/10.48550/arXiv.1301.3781.

Nissim, Malvina, Rik van Noord, and Rob van der Goot. 2019. "Fair Is Better than Sensational: Man Is to Doctor as Woman Is to Doctor." arXiv preprint, arXiv:1905.09866v2 [cs.CL]. https://doi.org/10.48550/arXiv.1905.09866.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

Puddifoot, Katherine. 2017. "Dissolving the Epistemic/Ethical Dilemma over Implicit Bias." *Philosophical Explorations* 20 (Supplement 1): S73–S93. https://doi.org/10.1080/13869795.2017.1287295.

Roth, Lorna. 2009. "Looking at Shirley, the Ultimate Norm: Colour Balance, Image Technologies, and Cognitive Equity." *Canadian Journal of Communication* 34, no. 1 (March): 111–36. https://doi.org/10.22230/cjc.2009v34n1a2196.

Schluter, Natalie. 2018. "The Word Analogy Testing Caveat." *Proceedings of the 2018 Conference of the North American Chapter of the Association for Compu-*

tational Linguistics: Human Language Technologies, Volume 2, edited by Marilyn Walker, Heng Ji, and Amanda Stent, 242–46. Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-2039.

Smith, Angela M. 2008. "Control, Responsibility, and Moral Assessment." *Philosophical Studies* 138, no. 3 (April): 367–92. https://doi.org/10.1007/s11098-006-9048-x.

Smith, David. 2013. "'Racism' of Early Colour Photography Explored in Art Exhibition." *Guardian*, January 25, 2013. http://www.theguardian.com/artanddesign/2013/jan/25/racism-colour-photography-exhibition.

Suresh, Harini, and John V. Guttag. 2020. "A Framework for Understanding Unintended Consequences of Machine Learning." arXiv preprint, February 17, 2020, revision, arXiv:1901.10002v3 [cs.LG]. https://doi.org/10.48550/arXiv.1901.10002.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, et al. 2016. "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation." arXiv preprint, arXiv:1609.08144 [cs.CL]. https://doi.org/10.48550/arXiv.1609.08144.

Oisín Deery is an assistant professor of philosophy at York University, Toronto, Canada, and a lecturer and ARC DECRA Fellow at Macquarie University in Sydney, Australia. He has previously held positions at Monash University, Florida State University, the University of Arizona, and the University of Montreal. Oisín's work straddles the philosophy of mind and action, moral psychology, and metaphysics. His work has focused primarily on developing a naturalistic understanding of human agency and responsibility, but recently he has also been working on issues related to artificial intelligence and agency. Oisín has published widely in journals such as *Philosophical Studies*, the *Australasian Journal of Philosophy*, and *Philosophical Psychology*. He is the author of a monograph entitled *Naturally Free Action*, published by Oxford University Press in 2021.

Katherine Bailey has worked in a number of roles as a data scientist and software engineer. She is currently employed by Shopify Inc., in Toronto, Canada. Katherine has a longstanding interest in issues related to artificial intelligence and natural language processing. She has written on these issues in blog posts and in publications such as *TechCrunch* and has been invited to speak at various venues, including Google and the North American Association for Computational Linguistics.