Twitter Response to "Proposal for Identifying and Managing Bias in Artificial Intelligence"

At Twitter, we believe the journey to responsible, responsive, and community-driven machine learning (ML) systems is a collaborative one. Responsible technological use includes studying the effects ML tools can have over time. When Twitter uses ML, it can impact hundreds of millions of Tweets per day and sometimes, a system can start to behave differently than was intended. These subtle shifts can then start to impact the people using Twitter and we want to make sure we're studying those changes and using what we learn to build a better product.

As part of our efforts to ensure responsible use of this technology, Twitter launched a company-wide initiative called Responsible ML, centered on ensuring equity and fairness in outcomes, promoting transparency about our decisions and how we arrive at them, and enabling algorithmic choice. We applaud efforts by NIST to further examine and provide guidance on the responsible development and use of ML systems.

Twitter has submitted specific textual recommendations to the framework. In addition, we write to highlight specific recommendations related to the approach the framework takes with regards to bias.

Recommendation 1: Reframe the guidance to address impacts of AI more broadly, not just bias.
In recent years, much of the discussion about AI has placed "bias" as the primary problem to be uncovered and mitigated. In the context of AI, the term "bias" can have multiple meanings. This definitional ambiguity has led to unproductive arguments in which the core points of contention are obfuscated by technical arguments about what does or does not constitute bias. Furthermore, there are many ways in which AI can be harmful to individuals or society that are not—under any reasonable definition—bias. By focusing solely on bias, we limit ourselves to focusing on one subset of the broader set of issues one ought to be concerned about when it comes to preventing AI harms and promoting positive social impacts.

People often think of bias as a "tendency, feeling or opinion, especially one that is preconceived or unreasoned."[1] Although this is related to AI in that such human biases may influence or be directly encoded in the data that serves as the foundation to AI systems, this is typically not the type of bias at issue when discussing AI-based systems. When it comes to AI, statistical bias and societal bias are typically what's being discussed.

[1] Based on the dictionary.com definition.

Statistical bias is defined in this report (and adopted from the International Organization for Standardization) as "the degree to which a reference value deviates from the truth." We interpret this definition to be fairly general, covering any measure of deviation between some model-relevant value and the "truth" it is meant to represent.[2] In the context of AI, evaluating statistical bias asks us to compare the outputs of the model (or the data on which it's built) to some ground truth. Truth, here, refers to *the world as it is*.

Statistical bias is considered particularly problematic when the degree of deviation between the model's output and the "truth" systematically differs across individuals or culturally relevant categories. For example, when model performance metrics vary differentially along social axes (like race), a model is often described as biased with respect to that axis, e.g. "racially biased." These types of model performance disparities have attracted the most attention over the past several years.

Societal bias occurs when *the way the world is* is itself the product of unjust or unfair processes. Like statistical bias, societal bias is a relative concept. Where statistical bias is defined relative to the world as it is, societal bias asks us to compare the outputs of a model (or the data on which it's built) to *the world as it should or could be.*

When it comes to AI, technical perspectives and definitions tend to dominate. In the case of "bias" this means that statistical bias is often over-emphasized. This definitional ambiguity opens the door for the type of technical fixes that addresses statistical bias but fails to address or even acknowledge societal bias. Even describing the task as mitigating "harmful bias" in practice renders this document ineffectual. This framing invites the response that because statistical bias has been eliminated (or did not exist in the first place), any harmful aspects of the model are not—in the technical sense of the word—"bias" and so any guidelines based on this definition do not apply.

We have more or less seen this play out in the now canonical example of racial bias in criminal justice risk assessment models. In summary, a model that predicts whether a person will recidivate was found to exhibit racial disparities in some model performance metrics. Following this initial claim of racial bias, proponents of such models scrambled to demonstrate that the racial disparities uncovered in the initial report were not indicative of "bias" because they merely reflect the unfortunate reality that Black people are arrested more than white people.

Despite the technical acumen on display in arguing for the lack of statistical bias in such models, these sorts of analyses have done little to assuage the concerns of risk assessment opponents. For example, a statement of concern released by several leading civil rights organizations following the back and forth about statistical bias warned of the potential for such tools to exacerbate existing racial disparities and called for tools to only be used to "reduce and ultimately eliminate unwarranted racial disparities across the criminal justice system." Statistical facts about current arrest patterns were not in their view enough to justify the racial disparities reproduced by the models. Reading between the

[2] For a detailed breakdown of the many ways the word bias is used in a statistical or machine learning context, see the technical appendix.

lines, this is because criminal justice advocates are comparing the model to their idea of the way the world should be—one with reduced or nonexistent racial disparities. They are talking about societal bias. The researchers involved in building the tools were comparing its performance to the way the world is. They were talking about statistical bias.

This dual definition of bias has led to largely unproductive conversations in which both sides talk past each other—one side convinced it has dispelled the existence of a problem that the other side was not fundamentally talking about in the first place. By framing directly around impacts, we can circumvent technical demonstrations illustrating that an AI-based system adequately reproduces the way the world is and talk directly about the world as we'd like it to be and whether the system moves us towards that goal. Definitional ambiguity aside, framing solely around bias explicitly excludes harmful impacts of AI that are equally applied across a population or for which there is no obvious baseline—based in reality or normative beliefs—with which to compare. For example, consider a highly addictive AI-based game in which all users are treated equally. Because there's no obvious single outcome predicted, typical bias evaluations that look for differential predictive performance would not apply. Furthermore, it is difficult to propose a normative target for time spent in the application against which to evaluate deviation from an ideal. How much time *should* a user spend in the application? Despite the lack of obvious "biases" to evaluate, addiction to the application could be harmful. This type of case is not covered by considering bias alone.

To address this we recommend focusing on impacts directly. There are many ways in which AI can have negative (and positive) impacts on people and society. Bias is only one mechanism by which harm arises. Framing AI harms solely around bias unnecessarily constrains mitigation efforts to a subset of potential solutions and limits our ability to see other types of harm. Furthermore, because of the primacy of technical solutions in technical domains like AI, this framing elevates technical analyses of statistical bias rather than inviting the more challenging and more human task of precisely defining the impacts on people and society we are trying to achieve by introducing an AI-based system. Inevitably, just as not everyone agrees about whether certain models are biased, not everyone will agree on desirable or fair individual outcomes or a vision for the society we want to live in. However, by focusing directly on impacts, we can avoid unproductive technical proxy wars about what is or is not "bias" and be sure we don't a priori exclude consideration of non-bias-based harms.

Recommendation 2: Address difficulties with measuring bias, including those that stem from practices designed to protect privacy.

Even if a measure of bias is agreed upon, it can be challenging to measure bias in practice. The current report fails to acknowledge this challenge.

First, most standard bias evaluation techniques require knowing demographic information, like race or gender. In many scenarios, this demographic information is unknown. For example, in the tech industry, demographic information is oftentimes not collected due to privacy and ethical considerations. In some settings like non-mortgage credit issuance, it is outright illegal for financial institutions in the United States to collect race, color, religion, national origin, or sex information. This

can lead to relying on proxies of demographic information (e.g. surname and geographic location as proxies to race and ethnicity), to predicting demographic information (which can be ethically problematic), or to relying on self-selected individuals who volunteer their demographic information. Each of these is subject to their own biases which can ultimately lead to an incorrect estimate of the bias of the original system.

Second, specifying the demographic groups with respect to which bias is evaluated requires great care. For example, it is insufficient to measure bias along each demographic dimension separately. In particular, it is possible an AI system does not appear biased with respect to groups defined by race or gender, but is biased to intersectional subgroups, e.g. Black women.[3] In some cases, discrimination law offers guidance on group definitions in the form of clearly defined "legally protected classes." However, even government standards can be insufficient. For example, using US Census definitions for racial groups can marginalize subgroups: Middle Eastern individuals are labeled as "white", and "Asian" encompasses a set of countries with varied cultures and socioeconomic statuses. Grouping Middle Eastern and Asian people in such broad categories can easily lead to insufficient measurement and undetected bias.

Finally, the technical focus on statistical bias makes an inherent assumption that we have access to high quality information on "the way the world is." In fact, the dataset collection and curation process is often rife with societal biases, excluding large swaths of the population that will be impacted by the model. There is no oracle to tell us the ground truth state of the world. Any attempt to minimize statistical bias alone is likely fitting to an incomplete and exclusionary target, amplifying existing societal biases in the process.

Recommendation 3: Require continuous monitoring of deployed models to mitigate harm.

While the document places an emphasis on vetting models before they are deployed, we must also consider the fact that deployed models are often continuously exposed to new data. Models can drift away from any baseline that was present when they were first deployed. The model itself often has an influence on the data it is collecting, leading to feedback loops in the system. For example, if a recommender system starts out recommending particular kinds of content, and finds that a user is interacting with that content, it will receive a positive signal that influences its future decisions when it is retrained. This kind of loop can be useful, but it also ignores the possibility that if a different type of content had been served in the first place, the model may have received a different feedback signal. Thus, models can drift into harmful territory even if they started out with a negligible amount of disparity because that initial disparity is amplified in the feedback loop. As a result, continuous monitoring of deployed models is necessary to ensure that harms are being mitigated.

[3] http://gendershades.org/

Technical Appendix

Definitions of statistical bias

In this note, we adopt a broad definition of “statistical bias” based on a liberal interpretation of the definition used in the NIST report. This may be confusing for readers who are familiar with other technical definitions of bias that are commonly used in statistics and machine learning. Here, we mention a few definitions that are often considered “statistical bias.”

Measurement bias occurs when the data systematically over- or under-represents the concept it is meant to measure. Differential measurement bias occurs when the degree of over- or under-representation is systematically different by group. Under this type of bias, the “truth” is the underlying concept of interest .

Sampling bias occurs when some elements of a target population are more or less likely to be represented in the data. Here, the “truth” is the target population, and deviations from that truth occur because the data over- or under-represents certain elements.

Estimation bias is the difference between an estimator's expected value and the true value of the parameter being estimated. Here, the truth is some “state of nature” and deviations from it are measured in terms of the expected difference between the true state of nature and the function of data we used to estimate it.

We note that the definition we adopt does not limit the measure of deviation from the truth to only differences in expectations as in the standard definition of bias in statistical estimation.