National Institute of Standards and Technology
Attn: Information Technology Laboratory
100 Bureau Drive
Gaithersburg, MD 20899

Via Email: ai-bias@list.nist.gov

RE: A Proposal for Identifying and Managing Bias within Artificial Intelligence (Spec. Pub. 1270)

New America's Open Technology Institute (OTI) submits these comments in response to the National Institute of Standards and Technology (NIST) June 2021 publication, "A Special Proposal for Identifying and Managing Bias within Artificial Intelligence" ("the proposal"). OTI has done extensive research on how internet platforms use artificial intelligence (AI) and machine learning (ML)-based tools[1] and on the risks posed by facial recognition systems.[2] We have also made recommendations on how companies and governments can promote greater FAT around the use of these algorithmic systems, particularly high-risk AI.[3] We appreciate NIST turning its focus to artificial intelligence and, in particular, the subject of how to mitigate bias in AI, which is a well-documented and growing issue. OTI recommends that NIST strengthen its proposal by further developing its guidance on transparency, pre-deployment assessments, post-deployment analysis, and use of proxy data and other data collection concerns.

## Executive Summary

We recommend that NIST take the following steps as it continues to investigate how to mitigate bias in AI:

1. Include transparency as one of the necessary characteristics for promoting trustworthy AI and mitigating bias in AI.
2. Create guidelines or encourage legislation that would  mandate testing of high-risk AI systems before deployment and create clear indicators for when a high-risk system should not be deployed at all.
3. Encourage developers of medium and high-risk AI systems to define intended use cases and identify situations in which their systems could generate harmful or inaccurate results. NIST should also push developers to provide a basic, public explanation of these systems along with access to an appeals process.
4. Advocate for all AI systems to be subject to continuous evaluation after being deployed.
5. Provide guidance on whether developers should use proxy data or collect and use sensitive demographic data when attempting to evaluate and mitigate bias.
6. Ensure that all guidance and recommendations consider the capabilities and needs of smaller entities developing AI systems.

[1] https://www.newamerica.org/oti/reports/report-series-content-shaping-modern-era/,

[2] https://www.newamerica.org/oti/briefs/civil-rights-concerns-regarding-law-enforcement-use-of-face-recognition-technology/

[3] https://www.newamerica.org/oti/reports/cracking-open-the-black-box/

## Recommendations

1. **Include transparency as one of the necessary characteristics for promoting trustworthy AI and mitigating bias in AI.**

While NIST has outlined numerous technical factors such as explainability and privacy that are integral for promoting trustworthy AI and mitigating bias, we believe transparency is also critical. As the report points out, most Americans are unaware when they are interacting with algorithmic systems. Transparency measures that are designed with users in mind can help address this issue.

We recommend that NIST include a reference to the need for transparency measures in any future proposals. If transparency is considered a component of one of the already mentioned categories, this should be clarified in the publication.

2. **Create guidelines or encourage legislation that would mandate testing of high-risk AI systems before deployment.**

As the report outlines, some technologies are not tested extensively or at all before deployment. Rather, developers use deployment scenarios to test their technologies. This can result in harmful and concerning outcomes. To avoid such outcomes, NIST should develop guidelines or encourage the creation of legislation that requires developers of high-risk algorithms to properly test their systems before they can be deployed.

NIST should also help create clear indicators around when a system can be green lighted to be deployed and when it cannot, within the context of a risk-based framework. Such a risk-based evaluation may also be helpful during the pre-design and design phases. If a system poses too significant of a risk to society and fundamental rights, it shouldn’t be deployed at all. Any efforts to promote FAT around this kind of system in the development, deployment, and post-deployment phases will be meaningless if the system is inherently high-risk. NIST should be clear that there may be some high-risk circumstances in which AI will never be safe enough to use, and no risk-mitigation will be enough. We define high risk-algorithms as systems that pose “high risks” to the fundamental rights and freedoms of citizens and society and medium-risk algorithms as systems that pose a moderate risk to the fundamental rights and freedoms of citizens and society.[4] We recognize that there are ongoing dialogues around defining high-risk algorithms. This is also an area where NIST could provide input and guidance.

3. **Encourage developers of medium and high-risk AI systems to define intended user cases and identify situations in which their systems could generate harmful or inaccurate results. NIST should also push developers to provide a basic, public explanation of these systems along with access to an appeals process.**

[4] https://www.newamerica.org/oti/reports/cracking-open-the-black-box/

We believe that NIST should recommend that developers of AI systems, particularly medium and high-risk AI systems, outline the intended use cases of their tools as well as cases in which the use of their systems could generate harmful or unreliable results. This is similar to the information encompassed in Model Cards.[5]

Additionally, we suggest that NIST recommend that developers and deployers of medium and high-risk AI systems provide a basic, public outline of how their algorithmic systems function to users. Providers of high-risk AI systems that have consequential impacts (e.g. credit algorithms) should also give users the ability to appeal decisions made by the systems. Providing users with access to an appeals process also helps expand user control and agency over systems that are often responsible for making critical life decisions. If providing an appeals process is not scalable, then developers should, at the least, enable users to understand what factors went into informing the decision.[6]

We believe these approaches can help promote transparency and accountability around harms that can result from certain AI systems. In some cases, this information could also help mitigate such harms. As previously noted, however, certain systems may be too high-risk to deploy. The design and deployment of such systems should not be permitted.

4. **Advocate for all AI systems to be subject to continuous evaluation after being deployed.**

NIST proposes addressing bias in the design, development, and deployment stages of AI systems. However, there is not as much emphasis on continuing these practices post-deployment. A pre-deployment evaluation of a system may indicate that the system is low-risk. But, AI and ML-based systems are constantly changing, learning, and adapting. Additionally, a system may be deployed in a new context. Both of these factors can change the risk potential of the system.[7]

NIST should add a fourth stage, post-deployment, to their framework for evaluating AI systems. NIST should also mandate that AI-producing entities continue to address bias in post-deployment scenarios. For example, if a developer conducts a risk assessment or bias evaluation during pre-deployment, they should continue to conduct such assessments post-deployment, particularly if the AI system is being used in a new context or if it has changed in some way.

5. **Provide guidance on whether developers should use proxy data or collect and use sensitive demographic data when attempting to evaluate and mitigate bias.**

As NIST points out, many companies use proxy data to inform their algorithmic systems. However, proxy-based inferences are not always accurate and can result in biased and

[5] https://arxiv.org/abs/1810.03993

[6] https://www.newamerica.org/oti/reports/cracking-open-the-black-box/

[7] https://www.newamerica.org/oti/reports/cracking-open-the-black-box/

discriminatory outcomes.[8] The proposed alternative to using proxy data is to have companies collect demographic data such as gender and race from users. However, there is little trust in algorithmic systems and companies in certain industries who are deploying them (e.g. internet platforms). There are also few safeguards to protect the collection and use of this data (e.g. the U.S.does not have comprehensive privacy legislation).[9] As a result, we do not believe collective sensitive demographic data is an appropriate solution as it could create new harms and exacerbate existing harms caused by algorithmic systems. However, some researchers and civil rights groups have pushed for the collection of racial data as it could enable audits of racial discrimination to take place. It would be helpful for NIST to provide guidance on how to strike a balance between using proxy data and the collection of demographic data. For example, in what situations, if any, should entities collect race-based data? When does the privacy interest in preventing the collection of race-based data outweigh the collection and use for countering bias in AI? What safeguards should exist around the collection, use, and storage of this data? What kind of agency and control do users have over the collection of this data? In what situations, if any, could proxy data be useful for race-based evaluations?

6. **Ensure that all guidance and recommendations consider the differing capability and needs of smaller entities developing AI systems.**

The report discusses how multistakeholder and interdisciplinary experts can help AI developers identify and mitigate harmful outcomes. While NIST recognizes that setting these kinds of processes up requires deliberate planning and guidance, it does not consider that larger companies and deployers may have greater access to these resources than smaller ones. Because of this, smaller entities may be at a disadvantage. NIST should provide guidance to smaller entities seeking to obtain feedback from multistakeholder and interdisciplinary experts, and ensure that they have the resources necessary for proper oversight and assessment of their AI systems.

Thank you for taking the time to consider our recommendations and concerns.

Respectfully,

Spandana Singh
Open Technology Institute
New America
740 15th Street NW
Washington, D.C. 20005

[8] https://www.newamerica.org/oti/reports/special-delivery/
[9] https://www.newamerica.org/oti/reports/automated-intrusion-systemic-discrimination/.