CENTER FOR
DEMOCRACY
& TECHNOLOGY

September 10, 2021

National Institute for Standards and Technology
100 Bureau Drive (Mail Stop 8940)
Gaithersburg, Maryland 20899-2000

Re: Proposal for Identifying and Managing Bias in Artificial Intelligence (SP 1270)

To: Reva Schwartz

In response to NIST's call for feedback on "Proposal for Identifying and Managing Bias in Artificial Intelligence" (Special Publication 1270), the Center for Democracy & Technology would like to make the following comments:

1. Context-agnostic approach to AI

Section 1. Lines 223-228; Section 2. Lines 262-264.

Comment

The proposal takes a universal approach to bias in AI and points out the limitations of approaches that "classify bias by type (i.e.: statistical, cognitive), or use case and industrial sector (i.e.: hiring, health care, etc.)" Although in places the document acknowledges the variety of impacts biased AI can have depending on the complexities and dynamics of a given context, it does not adequately address the limitations of a universal, context-agnostic approach to bias in AI, particularly in contexts where human rights and anti-discrimination laws are significantly at issue, such as financial services or the criminal justice system (see comments from the PASCO Coalition for an example of the dangers of deploying AI without considering the human rights implications).

Suggested change

NIST should be more specific about the role and limitations of a universal approach to bias in AI. It should recommend that the document be used in conjunction with more actionable, context-specific approaches to bias in AI. Additionally, NIST should provide some guidance on situations where the use of biased AI could raise human rights concerns or violate anti-discrimination laws. While it is likely not within NIST's purview to provide comprehensive information about which contexts will raise heightened concerns, NIST should add examples of such contexts to the document, along with information about the particular damage a biased AI system could cause in that context. Ideally, this would be accompanied by a framework to help developers evaluate their own context to determine the potential harms their systems could cause. This discussion is related to the issue of repurposing AI systems across contexts (lines 630-655), which we commend NIST for including, and encourage it to build on that discussion as a way to highlight the context-specific elements of bias that can plague AI systems.

2. Lack of definitions

Section 1.

Comment

The proposal lacks clear definitions for fundamental terms such as “risk”, “harm”, and “artificial intelligence system”. NIST may want to leave room for context-specific authorities to define such terms themselves. However, readers may not have sufficient guidance to come up with their own definitions.

Suggested change

Should NIST choose not to define these terms themselves, it should explicitly delegate responsibility for such definitional work to context-specific authorities. NIST should provide guidance on how to define these terms, either with a separate framework or through a broader range of examples of bias in AI and examples of risk or harm.

3. Population-level harms

Section 2. Lines 239-252.

Comment

When discussing the challenges posed by bias in AI, the proposal gives use cases where AI systems can cause significant harm to individuals, such as in hiring, health care, and criminal justice. It does not include examples of AI systems that could harm entire populations, such as recommender systems and content moderation algorithms on social networks with hundreds of millions or billions of users. Such harms are not always as salient or easily observable as individual harms, but the sheer scale means they can have a significant impact.

Suggested Change

The proposal should articulate a view of risk that considers both the severity and the scale of harm an AI system poses. The proposal should explicitly call out this latter mode of risk and include references to the literature on bias in content moderation and recommender systems on social networks (for example, CDT’s “Do You See What I See” and “Mixed Messages?” reports). NIST should also use these examples to highlight human rights issues raised by bias AI, as explained in Comment #1.

4. Encourage means of redress

Section 2. Lines 242-243.

Comment

The proposal suggests that adopting a risk-based framework encourages technologists to develop means of “managing and reducing the impacts of harmful biases in AI”. The document fails to mention that AI systems should have built-in means of redress for when bias does occur.

Suggested change

NIST should recommend that all AI systems provide channels for dispute and remedy, and that these channels be designed to be robust against abuse. The proposal should recommend that AI systems include a means of redress regardless of a system’s supposed riskiness or the severity of its potential harm. Redress should be more robust when the use of AI creates a risk of substantial harm to an individual and when it creates a risk of less salient harms to many individuals. Dispute and redress systems should be designed in such a way that developers can learn about potential systematic errors or biases in the system based on the complaints and feedback they receive. In order to achieve this, users must be aware of, and have easy access to, any redress mechanisms. Additionally, redress systems need to be sufficiently transparent and understandable to users that they can meaningfully understand the action taken by the system and, if necessary, explain why they feel that action is incorrect or harmful.

5. Role of audits

Section 3. Lines 387-389.

Comment

The document positions audits as the primary tool for limiting bias in AI systems. Audits can be a helpful tool for accountability and improvement, but they should not be considered sufficient without other forms of transparency about a given AI system. Focusing on audits without also providing for other forms of transparency overlooks the role of, among others, journalists, academics, impacted communities, and civil society in raising and addressing these issues.

Suggested change

NIST should position transparency about the design, inner workings, and real-world consequences of AI systems as a necessary companion to auditing. NIST should discuss a range of transparency mechanisms that can raise public awareness of harmful AI bias, including transparency reports, researcher data access programs, and user notifications about when an AI system is used to make a decision.

When discussing audits, NIST should recommend that they occur during both the development and the deployment stages of the AI lifecycle. Audits in the development stage allow developers to catch errors before their systems are released into the wild; audits after deployment allow auditors to take action when an AI has real-world consequences. Auditors need more than expertise to be effective. They also must be independent and document their findings, and there should be a defined mechanism through which the bias that they flag is addressed and mitigated.

6. Explainability vs trusted authorities

Section 3. Lines 387-390.

Comment

The proposal suggests that the way to build public trust in AI is through empowering an authority of expert AI auditors rather than through building trust in individual AI systems. This is an unduly cramped and incorrect conclusion. Many communities, particularly marginalized communities that are most likely to be affected by bias, may not trust alleged experts and in any case will be guided in their views by their lived experience with AI systems. Moreover, members of impacted communities, particularly those that are already marginalized or at-risk, may be more aware of and adept at detecting bias than auditors. Explainability and transparency are crucial to building trust -- a “black box” system simply will not lead to trust even if it is “validated” by an auditor.

Suggested change

AI systems should be designed in a way that lets experts and, critically, affected communities, evaluate algorithmic outputs for themselves. NIST should recommend that AI systems offer explanations for how they come up with decisions in order to let experts and communities better evaluate them. We recognize that explainability may be covered in future NIST proposals, but it is critical to recognize the roles of explainability in engendering trust and providing a means for detecting and managing bias. Additionally, we recommend that any future work on explainability also discusses the importance of access for marginalized communities, such as those who may not speak the same language as system designers.

7. Addressing bias when it is discovered

Section 4. Lines 418-422.

Comment

The proposal provides a framework to identify bias in an AI system, but does not offer guidance on how to address bias once discovered. Notwithstanding its title, the proposal is significantly skewed towards "identifying" bias in AI rather than "managing" it.

Suggested change

The proposal would be more useful for developers and more effective in improving the overall state of AI if it were able to provide guidance and resources for mitigating bias once discovered, including steps for identifying and evaluating potential solutions. If the bias in a given AI system is impossible to mitigate, we encourage NIST to recommend that this system not be deployed. We recognize that bias mitigation may be beyond the scope of this document in some cases, in which case we recommend the document provide references to existing guidance from other sources, such as the Brooking Institute's report on Algorithmic Bias Detection and Mitigation.

8. Post-deployment bias

Section 4. Lines 654-664.

Comment

The proposal implies that AI developers can sufficiently prevent their systems from producing biased behavior and outcomes by accounting for bias during the pre-design, design, and development stages. However, systems may exhibit unexpected behavior or results when placed in a live environment. Consequently, assessment done during the design and development process may be incomplete or incorrect.

Suggested change

The proposal should place more emphasis on post-deployment monitoring and evaluation. It should make clear that managing bias is an ongoing, iterative process that must be done throughout the AI lifecycle, rather than an evaluation done at a single point in time.

9. Commendation for acknowledging diversity

Section 4. Lines 475-478.

Comment

We commend NIST for the guidance to assemble and support diverse teams of designers and developers. We do feel the language could be more inclusive where disability is concerned.

Suggested change

We encourage NIST to ensure this point is present in the final version of the guidance. Additionally, we recommend the guidance use “disability diversity” in place of “diversity of physical ability,” as many disabilities may not fall under the umbrella of physical ability.

10. Commendation for acknowledging deception

Section 2. Lines 319-333.

Comment

We commend NIST for acknowledging the presence of “snake oil salesmen” in the field of AI and the importance of evaluating systems on their efficacy in addition to considerations like non-discrimination and safety.

Suggested change

No change is recommended, rather we encourage NIST to ensure this point is present in the final version of the guidance.

Sincerely,

Gabriel Nicholas
Research Fellow, CDT

Hannah Quay-de la Vallee
Senior Technologist, CDT