September 10, 2021

Via electronic mail
National Institute of Standards and Technology
Attn: Information Technology Laboratory
100 Bureau Drive
Gaithersburg, Maryland 20899-2000 70
Email: ai-bias@list.nist.gov

Re: Request for Comment on SP1270-Draft *A Proposal for Identifying and Managing Bias within Artificial Intelligence*

Dear NIST Information Technology Laboratory Team,

We, the undersigned group of public defenders and advocates who focus on forensic and data science, submit these comments in response and opposition to the *A Proposal for Identifying and Managing Bias within Artificial Intelligence* draft, NIST SP1270-DRAFT (hereinafter, "Draft Proposal"), published for comment in June 2021.

As set forth in detail below, we strongly oppose publication of the Draft Proposal. We urge NIST to continue grappling with and exploring the critical impact of bias in systems of artificial intelligence. However, work in this area must start from first principles:

> *"The only remedy to racist discrimination is antiracist discrimination. The only remedy to past discrimination is present discrimination. The only remedy to present discrimination is future discrimination. As President Lyndon B. Johnson said in 1965, 'You do not take a person who, for years, has been hobbled by chains and liberate him, bring him up to the starting line of a race and then say, 'You are free to compete with all the others,' and still justly believe that you have been completely fair.* [1]*'"*

NIST's Draft Proposal suffers from a failure to grapple with these first principles as well as the underlying assumptions that the Draft Proposal thus necessarily makes. Most critically, the Draft Proposal assumes without scrutiny that the opposite of bias is mere fairness. But if the decades of struggle within the criminal legal system have taught us anything it is that the opposite of bias is not just fairness, but an intentional move towards justice.

[1] Ibram X. Kendi, *How to be an Antiracist*. (New York: One World, 2019).

Justice emanates not merely from the product, but also from the process. In multiple places SP1270 acknowledges the "benefit of engaging a variety of stakeholders and maintaining diversity along social lines where bias is a concern (racial diversity, gender diversity, age diversity, diversity of physical abilities)." And correctly identifies part of this benefit as "lead[ing] to a more thorough evaluation of the broad societal impacts . . ." However, the SP1270 authorship team does not reflect this principle. Furthermore, SP1270 does not even affirmatively indicate alignment with this principle through the use of an advisory committee or other mechanism. This failure is particularly galling given that there is a robust community, literature and practice to draw from. NIST's failure to do so here is indefensible.

There is no way around bad process. As a federal agency, NIST should truly engage in anti-bias work. Doing so would require holding publication of SP1270, reconstituting the authorship team to include those most directly affected by AI bias, and prioritizing opportunities for equitable partnership. In the realm of addressing bias, at a minimum, commitment to an anti-racist lens is non-negotiable.

Justice and fairness are different principles. SP1270 states: "The goal is not 'zero risk,' but to manage and reduce bias in a way that contributes to more equitable outcomes that engender public trust."

This is a breathtaking assertion. Echoing Justice Brennan almost thirty-five years ago:

> *"Taken on its face, such a statement seems to suggest a fear of too much justice. Yet surely [the authors] would acknowledge that if striking evidence indicated that other minority groups, or women, or even persons with blond hair, were disproportionately [affected], such a state of affairs would be repugnant to deeply rooted conceptions of fairness."*[2]

Zero risk *has to be* the goal, even if it is an unattainable one. For a federal agency to state otherwise communicates (once again) that Black, Latinx, and brown communities are deserving of less justice.

NIST should clearly state that *the goal* is zero risk. And, in light of this reframing, SP1270 should reject the entire premise of "responsible AI." The "responsible AI" construct makes it difficult, if not impossible, to question AI deployment decisions. The baseline question must shift from "can we," to "should we?" because the most important implication of a conclusion that risk is unavoidable is one that SP1270 does not acknowledge.

[2] *McCleskey v. Kemp*, 481 U.S. 279, 339 (1987) (Brennan, J., dissenting).

That is: what do we do in situations where the risk of harm cannot be reduced to a non-impactful level? Are there some use cases that must be off limits? What guidance can NIST offer about when *not to* deploy an AI system? The answers to these questions cannot be silence. Just as some treatments for disease pose such unacceptable risks to the public that they cannot be approved for use, some AI projects are similarly incapable of just application.

Justice requires rejecting the surveillance mandate. Every AI system in production, by definition, must have some data streams to draw from. These data streams are what AI makes predictions about in use and they are also what AI systems use for validation and correction of predictions. Therefore, every AI system has a mandate for surveillance to some degree.

SP1270 appears to assume that the data (which constitute the "raw materials" from which AI is built) is already in the hands of the developer. This means that SP1270 does not adequately anticipate an additional hazard emanating from a bias management framework: if "reducing bias" requires collecting more information – e.g. broadening the scope of surveillance – then one cannot assume straightforwardly that making AI *fairer* will result in less harm overall.

This is far from a speculative concern. Recently details of a project of the Department of Homeland Security's Homeland Security Investigations became public. The data/analytics platform, called RAVEn, is being proposed to ingest data from a diverse set of sources and make it searchable by ICE. In the Privacy Impact Assessment for RAVEn, DHI states the following:

> *"Pattern isolation is most successful if a tool has all relevant information and large datasets, thus the more information ingested by the tool will dramatically decreases [sic] the risk of introducing error or bias into RAVEn machine learning models."*[3]

In other words, under DHS's reasoning here, increased data ingestion and surveillance (i.e. very real and detailed information about actual human beings) will decrease bias. If the mandate of reducing bias can be claimed by organizations as a rationale for increasing their surveillance powers, then a framework for bias reduction will not serve as a constraint on the dangerous uses of AI that SP1270 assumes it will.

[3] Dep't of Homeland Security, *Privacy Impact Assessment for the Repository for Analytics in a Virtualized Environment*, DHS/ICE/PIA-055 (May 13, 2020), https://www.documentcloud.org/documents/21052700-privacy-pia-ice055-raven-may2020, at 8.

SP1270 fails to grapple with the problem of repurposed data. Taking the reasoning of DHS mentioned above at face value, the possession of "all relevant information" leads directly to the problem of repurposing, an issue largely unaddressed in SP1270. Specifically, the problem of repurposing arises when the data used to train one type of AI gets deployed against another problem via tinkering or experimentation. Once an agency or developer gets their hands on the data, they can easily expand the range of goals pursued with that data, without having to follow the process of review outlined in SP1270.

SP1270 must incorporate more straightforward ways of reducing bias: decommissioning or reducing the use of AI and the data collection methods that underpin it.

Overall, SP1270 suffers from an authoritarian tone. The Draft Proposal focuses uncritically and without support on "advancing AI" and "cultivating trust in AI systems." SP1270 seems to view the problem of bias in AI as a problem of perception (i.e. "how can we make AI look trustworthy to the public?"), rather than a problem of substance.

An example of this can be found at lines 385-90:

> *"A consistent finding in the literature is the notion that trust can improve if the public is able to interrogate systems and engage with them in a more transparent manner. Yet, in their article on public trust in AI, Knowles and Richards state '. . . members of the public do not need to trust individual AIs at all; what they need instead is the sanction of authority provided by suitably expert auditors that AI can be trusted.' Creating such an authority requires standard practices, metrices, and norms. NIST has experience…"*

This tone, itself, undermines SP1270's stated objective.

SP1270 should be edited to address first principles: "what human decisions will a given AI solution supplant, and why should we replace them?" Unlike, for example, a uniform system of measurements for industry, "artificial intelligence" is too broad a tent to uncritically promote advancement of. NIST needs to address the real implications of bias in AI and speak directly to the rise of the "New Jim Code,"[4] as well as the threat of automation bias and the need (in some instances) to use antiracist discrimination to correct past racist discrimination.

SP1270 fails to grapple with the overarching problem of a lack of data transparency. Multiple scholars and advocates have emphasized that secrecy in

[4] Dr. Ruha Benjamin coined the term "The New Jim Code," and defines it as: "the employment of new technologies that reflect and reproduce existing inequities but that are prompted and perceived as more objective or progressive than the discriminatory systems of a previous era." Benjamin, *Race After Technology: Abolitionist Tools for the New Jim Code* (Medford, MA: Polity, 2019).

development, procurement, implementation, and oversight directly impacts public trust. While “transparency” is mentioned 7 times in the report, every mention is found in the references list; SP1270 does not itself mention “data transparency” once.

A framework that fails to acknowledge the need for data transparency fails in its project of evaluating bias and fails in its project of building public trust. How can a member of the public trust a framework that offers no guidance on how failure and risk should be disclosed? SP1270 should reevaluate its framework to directly define the lack of transparency in the AI space (particularly in governmental use of AI solutions), and to include standards and guidance on the need for transparency to even begin to “identify” bias.

SP1270’s bias toward automated decision-making ignores the multiplicative risks of AI. Even when we assume *maximal* commitment to anti-racism and unbiased AI, additional risks to both justice and fairness remain. As Bainbridge [1983] states:

> *“The second problem is that if the decisions can be fully specified then a computer can make them more quickly, taking into account more dimensions and using more accurately specified criteria than a human operator can. There is therefore no way in which the human operator can check in real-time that the computer is following its rules correctly. One can therefore expect the operator to monitor the computer’s decisions at some meta-level, to decide whether the computer’s decisions are ‘acceptable.’ If the computer is being used to make the decisions because human judgement and intuitive reasoning are not adequate in this context, then which of the decisions is to be accepted? The human monitor has been given an impossible task.”*[5]

The impossibility of this task becomes entirely unmanageable where bias is present.

In unforeseen situations – those unanticipated by an AI system’s designers or engineers, or unrepresented in the underlying training data – human operators must step in to intervene, making a complex socio-technical judgement about how their system contributes to inequality or harm. However, due to the background inequalities of the society AI is deployed in, the risk of harm for the operators and maintainers will likely be far lower than for those affected by the AI system, creating both a technical and a moral hazard.

This reality underscores the need for both a representative design process and a meaningful and representative feedback mechanism for responding to risks that arise in deployment.

[5] Lisanne Bainbridge, “Ironies of Automation”, *Automatica* 19:6

But representation alone cannot fully solve automation's risk. Despite the fact that literature on the perils of automation has been around for decades, SP1270 pays little more than lip service to it.

Because the Draft Proposal fails to adequately address fundamental issues of justice, including process, privacy, and transparency, and does not embody anti-racist principles, the Draft Proposal should not be finalized. Instead, the legal and data science communities implore NIST to address these critical shortcomings and improve the Draft Proposal – both in process and in substance – prior to publication.

Sincerely,

Elizabeth Daniel Vasquez
Director
Science & Surveillance Project
Brooklyn Defender Services

Andrew Foltz-Morrison
Data Scientist
Science & Surveillance Project
Brooklyn Defender Services

Joseph Cavise
Forensic Science Division
Law Office of the Cook County Public Defender

Richard Gutierrez
Forensic Science Division
Law Office of the Cook County Public Defender

Kate Judson
Executive Director
Center for Integrity in Forensic Sciences

Julia Leighton
Retired
Public Defender Service for the District of Columbia

Janis Puracal
Executive Director
Forensic Justice Project

Emily Prokesch
former Forensic Practice Director
Bronx Defenders

Jessica Willis
Special Counsel to the Director on Forensic Science
The Public Defender Service for the District of Columbia