PEAT Response to NIST SP-1270

September 10, 2021

Introduction

On June 22, 2021, the National Institute of Standards and Technology (NIST) outlined their approach to addressing bias in artificial intelligence (AI) by publishing *A Proposal for Identifying and Managing Bias in Artificial Intelligence* (NIST Special Publication 1270). This publication is part of the agency's broader effort to support the development of trustworthy and responsible AI.

The Partnership on Employment & Accessible Technology (PEAT) has prepared this document to summarize feedback on the proposal. We have included general comments and line-specific edits to ensure NIST's framework can effectively help organizations identify and manage bias in their use of AI-enabled systems and processes. In addition, our mission is to ensure that all AI tools operate in line with principles of equity and accessibility, including for people with disabilities. We welcome further dialogue with NIST as the framework continues to take shape. You can reach out to PEAT by sending an email to: hello@peatworks.org.

General Comment: Identifying vs. managing bias and establishing priorities

A Proposal for Identifying and Managing Bias in Artificial Intelligence appears to treat the process of identifying and managing bias as a single concept. However, identifying bias and managing bias are different components. The paper refers to identifying and managing bias independent of context and across all contexts [line 228]. We argue that although guidance for identifying bias can be largely independent of context, guidance for managing bias is very dependent on context. In managing bias, organizations need to understand potential effects and prioritize solutions within the context of an AI application. Different types of bias are higher priority for management in different contexts, given that the severity of societal impact varies in different contexts.

In the proposed framework, there seems to be a disconnect between the pre-design stage of problem formulation, which focuses more on subject matter experts [lines 473-475, 494-495], and the design and development stage, which jumps directly to modeling and engineering with tech experts [lines 512-513]. What seems to be missing is a "solution formulation" or "solution architecture" stage, uniting both subject matter experts and tech experts. In this stage, there are some "solution framing questions" that we would like to suggest for consideration to be listed in a "Practices" section, including three distinct categories of contexts where bias identification and management in AI may

PEAT is funded by the Office of Disability Employment Policy, U.S. Department of Labor and managed by Wheelhouse Group PEAT material does not necessarily reflect the views or policies of the Office of Disability Employment Policy, U.S. Department of Labor, nor does the mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

be approached differently: (1) AI "Usage" Context; (2) AI "Domain" Context; and (3) AI "Impact" Context. These guiding questions are worth acknowledging as a key practice for helping to identify types of bias and manage according to prioritized severity. Each are described below.

AI Usage Context: Will my AI tool be used for assistance or for assessment?

We believe there is a distinction between AI-enabled "assistance", or AI tools that help people to do a job, and AI-enabled "assessment", or AI tools that make automated decisions about people. These need different considerations for identifying types of bias in the definitions section. For example, tools that provide (or lack) AI-enabled assistance (e.g., automated captioning, wayfinding via computer vision, etc.) may affect user metrics data that can lead to "activity bias." Tools that provide AI-enabled assessment, as in the real-world example of facial assessment referenced in lines 494-498, may lead to more harmful outcomes. Bias in the way a tool is used for AI-enabled assessment may be a higher priority for management.

AI Domain Context: What technical metrics affecting bias management are high priority in my domain?

Though the paper's aim is to guide bias identification and management across contexts, it is important to emphasize domain-specific priorities for addressing bias. Lines 257-259 comment that the rapidly lengthening list of contexts makes it difficult to develop overarching guidelines but could benefit from examples in this section. Between the pre-design and design/development stage, subject matter experts should pair with developers to determine a priori the most important technical metrics and acceptable concessions for downstream management of societal impact. For example, in medical diagnostics, a subject matter expert may deprioritize accuracy in favor of false positives while managing bias and accepting non-transparent black box models. Contrast this with a hiring context, where explainability may be prioritized and non-transparent black box models may not be acceptable because the variables leading to candidate selection are important for ensuring equitable outcomes. In a specific domain, technologists need to include subject matter experts for establishing a baseline comparison for benchmarking against human performance where bias perception may not meet reality (e.g., self-driving car accidents have high visibility for bias, but with a benchmark against human drivers the bias may be less/different).

AI Impact Context: Who or what is being assisted, assessed, or otherwise impacted by my AI application?

It is important to consider who and what segments of the population an AI application will impact. Where AI will impact multiple segments of the population, bias tradeoffs that need to be made should consider any specific groups of people who have more severe existing systemic bias that may be compounded by an AI application. An example is the priority of focusing on "activity bias" in the case where people with disabilities will be impacted users—when training data is collected on a population including people with disabilities interacting with a system, which may look different from people without disabilities interacting with a system.

Line-Specific Feedback

Line 209: After the sentence ending in "... a public lack of trust", suggested adding a sentence: "The risk of bias can increase with AI systems that store private information, offer surveillance features, enable automated decisions, or fail to adequately support human oversight." [Critical Platform Studies Group, Tech Fairness Coalition, and the ACLU of Washington; Algorithmic Equity (AE) Toolkit]

Lines 283-286: Suggest rephrasing to include the underlined phrase and new reference: "For 'employment suitability,' an AI algorithm might rely on time in prior employment, previous pay levels, education level, participation in certain sports [115], gaps in work history (which might disadvantage candidates with disabilities) [NEW REFERENCE: Lydia X. Z. Brown, Ridhi Shetty, Michelle Richardson (Center for Democracy in Technology), "Report – Algorithm-driven Hiring Tools: Innovative Recruitment or Expedited Disability Discrimination?" 03-December-2020.], or distance from the employment site [51] (which might disadvantage 286 candidates from certain neighborhoods)."

Line 286: Suggest adding another example specific to disability bias: "Machine learning algorithms can discover subtle correlations and proxies for protected characteristics like disability status, even when they are purposefully omitted from the model-building process. For example, a preference for large fonts could serve as a proxy for visual impairment or use of video captions could be correlated with deafness. Household income, educational achievement, and many other variables can also be correlated with disability." [NEW REFERENCE: Shari Trewin, Sara Basson, Michael Muller, Stacy Branham, Jutta Treviranus, Daniel Gruen, Daniel Hebert, Natalia Lyckowski, Erich Manser; "Considerations for AI Fairness for People with Disabilities"; AI Matters, Volume 5, Issue 3, September 2019, pp 40–63]

Lines 302-303: Add "disability" to the phrase in parentheses: "(Federal laws and regulations have been established to prohibit discrimination based on grounds such as gender, age, and religion.)

Line 317: Add another example: "They may also produce unjust outcomes for people with disabilities when automated decisions leveraging historical data can reproduce patterns of systemic discrimination [NEW REFERENCE: The Leadership Conference Education Fund, "Civil Rights Principles for Hiring Assessment Technologies", July 2020.]

Line 426-438: Consider during the problem formulation phase if the proposed AI system might reflect inherent prejudices or faulty assumptions related to disability. Some potential ways to do this include: (1) Identify applicable established or emerging regulations, standards and guidelines that apply to the proposed AI system; (2) Outline a non-automated process that would be automated by the proposed AI system, then document in the non-automated process where inherent prejudices or assumptions can be made that could perpetuate harmful biases; (3) Outline how the proposed AI system would automated the non-automated process and consider ways in which the inherent prejudices or assumptions could perpetuate bias; (4) Review the proposed AI system with internal AI Ethics / Equity Boards, DEI, Accessibility Programs, Legal to determine if the proposed AI system might perpetuate biases harmful to people with disabilities.

Lines 475-478: Expand "diversity of physical ability" to "Diversity of physical or cognitive ability" in the phase in parentheses (racial diversity, gender diversity, age diversity, diversity of physical or **cognitive** ability) [NEW REFERENCE: Shari Trewin (IBM), "AI Fairness for People with Disabilities: Point of View", 2019.].

Line 478: Recommend adding this practice after the sentence ending in "…diversity along social lines where bias is a concern (racial diversity, gender diversity, age diversity, diversity of physical ability) [32].":
"One way to identify unknown or potential negative impacts is to identify who might be excluded using practices such as 'Unknown Unknowns'" [NEW REFERENCES: (1) Kathy Baxter, "How to Build Ethics into AI — Part II", Medium, 02-April-2018; (2) Lakkaraju, Himabindu, Ece Kamar, Rich Caruana, and Eric Horvitz.

"Discovering Unknown Unknowns of Predictive Models", Paper presented at the 30th Annual Conference on Neural Information Processing Systems (NIPS), Workshop on Reliable Machine Learning in the Wild, Barcelona, Spain, December 9, 2016.]

Line 484: Consider adding this after the sentence ending in "…Technology or datasets that seem non-problematic to one group may be deemed disastrous by others.": "For example, consider how people with disabilities may be excluded or underrepresented in your dataset, why they are excluded and how to mitigate it. Teams can use practices like cost-sensitive learning, changes in sampling methods, and anomaly detection to deal with imbalanced classes in machine learning. They can also create different algorithms for different groups rather than one-size-fits-all or evaluate using path-specific counterfactual fairness. [NEW REFERENCE: Kathy Baxter, "How to Build Ethics into AI — Part II", Medium, 02-April-2018]

Lines 487-491: Just after the sentence ending in "… sometimes be identified in early testing stages, but are usually very specific to the contextual end-use and will change over time", add: "One way to mitigate evaluation bias is to ensure that the testing or external benchmark populations more equally represent the various parts of the user population who will use the proposed AI system, including a diverse mix of test subjects who have different racial, gender, and disability identities and lived experiences."

Line 491: Suggest adding another pre-design practice for accessibility: "During pre-design, teams should consider accessibility of the user interface (UI) for people with disabilities and conformance to established digital accessibility standards, so as to mitigate the risk of presentation bias. For example, if the proposed AI system has a conversational UI, teams should consider the needs of users with non-standard speech or who cannot or prefer not to speak by offering a non-speech alternative UI with equivalent functionality." [NEW REFERENCE: W3C Web Content Accessibility Guidelines]

Lines 493, 562: Both of these real-world examples reference bias in supervised learning algorithms. The paper could benefit from adding guidance with examples of bias in unsupervised learning.

Line 504-508: Downstream bias propagation can also result from bias in the way data are collected through interfaces that may not be accessible to all or require a group of people to disclose personal information to use a system (e.g., people with disabilities and requesting accommodation) even if those interfaces do not use AI.

Lines 495-498: Consider adding other examples related to facial recognition and people with disabilities: (1) "Facial recognition will be biased against people with craniofacial differences (e.g., Bilateral Microtia and Atresia, PRS, Treacher Collins Syndrome, Goldenhaar, hemifacial microsomia, etc.), and people who have had significant facial surgery." [NEW REFERENCE: Sheri Byrne-Haber, "Disability and AI Bias", 11-July-2019]; (2) "Facial and voice analysis technologies, in particular, have been shown to be inaccurate for people of color, English speakers with non-native accents, and transgender, nonbinary, and gender nonconforming people." [NEW REFERENCE: The Leadership Conference Education Fund, "Civil Rights Principles for Hiring Assessment Technologies", July 2020.]

Line 508: Add sentence with example of pre-design practices to ensure more inclusive representation of people with disabilities:
"For example, more inclusive representation of people of disabilities might involve determining if training data contain human decisions that are biased toward people with disabilities and are passing on institutional or systemic bias on to the learned model (e.g., college recruiters systematically overlook applications from students with disabilities, or a health insurer routinely denies coverage to people with disabilities, therefore a model trained on that data will replicate the same behavior." [NEW REFERENCE: Shari Trewin (IBM), "AI Fairness for People with Disabilities: Point of View", 2019]

Line 557: After "subject matter experts and practitioner end users" add the following: "who are gender, and disability identities and lived experiences".

Line 560: Suggest adding a new paragraph to make explicit the need to address **presentation bias** – i.e., potential biases arising from how information is presented to the end user via user interfaces (UI): "During the Design and Development phase, teams should work with disability advocates and accessibility consultants as well as utilize digital accessibility resources to reduce or mitigate potential presentation bias. This would include reviewing the design and testing developed code/technology to determine how they conform to applicable digital accessibility standards (e.g., W3C Web Content Accessibility Guidelines [NEW REFERENCE: W3C Web Content Accessibility Guidelines], recognized international standards that apply to applications and content, and which are incorporated by reference into accessibility laws and regulations). Teams should engage accessibility SMEs (e.g., internal accessibility programs, diversity/equity/inclusion employee resource groups, external consultants, etc.), and should perform accessibility testing with people who have various disabilities and who employ commonly used assistive technologies. Accessibility defects should be resolved prior to deployment, and process should be put in place to offer specific support for end users with disabilities such as offering reasonable accommodations or assisting with access to the AI-enabled system." [NEW REFERENCE: PEAT Staff Accessibility Training Resources]

Lines 601-606: After the sentence ending in "…likely creating downstream system activity that does not reflect the intended or real user population [1,8]", add:
"For example, if an algorithmic model is built on data only from the most active users, it might exclude people with disabilities who statistically may be less active than non-disabled users in the AI-enabled system."

Line 613: Consider adding example related to user interaction bias after the sentence ending in "… used in different settings and for different purposes, we see perceptions turn to unintended use cases and even distrust":
"In the deployed AI system, teams should determine if there points at which human users can impose their own biases about disability or disabled people into the system. Some practices to consider are doing a "pre-mortem" prior to deployment to consider ways in which the AI system could be abused and cause harm to disabled people [NEW REFERENCE: Kathy Baxter, "How to Build Ethics into AI — Part II: Research-based recommendations to keep humanity in AI", Medium, 2-April-2018] and setting up human-in-the-loop checks in place to prevent disability bias where possible. Once example could be: a recruiter using an AI enabled hiring technology and intentionally marking people who have requested reasonable accommodations as not qualified (just on account of their accommodation request) without determining if the candidates are able to perform the essential duties of the job."

Line 643: In addition to the disconnect between the groups who invent and produce a technology and those who use it, there can be a contextual gap during deployment due to technical ownership change and turnover of maintainers. An example approach to addressing this is an emphasis on documentation to prevent knowledge loss.

Line 653 Practical Improvements section: Consider emphasizing the importance of ensuring "reproducibility" in deployment as a practical improvement.

Line 653 Practical Improvements section: The example approach listed is a technical framework for monitoring and auditing. Consider re-emphasizing a human-in-the-loop approach for monitoring and auditing risk, which is briefly mentioned in line 579.