September 10, 2021

***VIA EMAIL***

National Institute of Standards and Technology
Att'n: Information Technology Laboratory
100 Bureau Drive
Gaithersburg, MD 20899
Email: ai-bias@list.nist.gov

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

ACLU
AMERICAN CIVIL LIBERTIES UNION

National Office
125 Broad Street
18th Floor
New York, NY 10004
aclu.org

**Deborah N. Archer**
President

**Anthony D. Romero**
Executive Director

We write in response to the National Institute of Standards and Technology ("NIST")'s June 2021 special publication "A Proposal for Identifying and Managing Bias within Artificial Intelligence" (the "publication").[1] We applaud NIST for seeking input on the critically important topic of artificial intelligence ("AI") and its potential for bias. The ACLU believes that the responses below will help inform NIST's policies and its effort to develop trustworthy AI as well as advance methods to understand and reduce harmful forms of AI bias.

## EXECUTIVE SUMMARY

In the draft publication, NIST has "identified the following technical characteristics needed to cultivate trust in AI systems: accuracy, explainability and interpretability, privacy, reliability, robustness, safety, and security (resilience)—and that harmful biases are mitigated."[2] Overall, the publication reflects an overly tech-determinist approach to mitigating bias in AI. While technical characteristics of AI must be addressed, numerous non-technical factors have contributed to the current need to improve trust in AI. These include, but are not limited to: the social context in which AI systems have been and will be used; the exclusion of communities and individuals that will be directly impacted by the use of AI systems from conversations about those systems' purpose, development, and deployment; the legal and regulatory context surrounding AI systems' use; organizational problems and practices that have led to a lack of public trust; failure to reform existing problematic AI systems and lax existing policy; and market incentives that encourage the use of untrustworthy AI. Additionally, the publication does not sufficiently emphasize lack of diversity among technical staff among the contributors of bias.

[1] Reva Schwartz et al., Nat'l Inst. of Standards & Tech., Spec. Pub. 1270, A Proposal for Identifying and Managing Bias in Artificial Intelligence (June 2021), https://doi.org/10.6028/NIST.SP.1270-draft [https://perma.cc/LM5M-ZD6Z].
[2] *Id.* at lines 105–07.

Biases in AI systems result directly and indirectly from both the technical characteristics of models and an array of non-technical factors. The ACLU has included within this comment recommendations addressing these technical characteristics as well as non-technical sociological and ethical considerations. As NIST undertakes the important challenge of identifying and managing bias in AI, we recommend that the following be addressed:

- Explicitly name impacted individuals and communities as key stakeholders in all aspects of the AI lifecycle.
- Meaningfully involve impacted communities in NIST's plans to develop a framework for trustworthy and responsible AI.
- Seek to improve institutions before pushing for additions of AI to these settings.
- Set standards for audits and impact assessments, including their public release and retention.
- Facilitate transparency about entities' uses of AI by, for example, producing research reports in partnership with others.
- Emphasize that algorithmic decision-making necessarily involves making policy decisions.
- Make clear that the economic costs of an algorithmic decision-making system include the costs of appropriate oversight, safeguards against bias, and compensation for individuals when misjudgment occurs.
- Revise the AI lifecycle to reflect the multiple points at which developers may decide to stop the development or use of an AI tool due to unacceptable biases and impacts.
- Develop a framework for assessing when the existence or risk of bias becomes intolerable in a given application.
- Highlight that data privacy violations, flawed or unethical collection methods, and data inaccuracy can all contribute to AI bias.

## I. Defining Stakeholders vs. Impacted Communities

**As a comment to:**

- Spec. Pub. 1270 lines 475–491

The draft publication makes broad references to stakeholders when discussing recommended community engagement, but this engagement appears limited to "experts" and "end users." While the publication recognizes "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 ability),"[3] no mention is made of impacted communities as stakeholders. Instead, the draft publication points to stakeholders such as "end-users, practitioners, subject matter experts, and interdisciplinary professionals from the law and social science."[4] Although the draft publication does acknowledge that AI tools may affect people other than end users, it does not center them in the AI lifecycle for managing biases. It is important that these impacted communities take part in the decision-making process at each stage, including the decision whether or not to adopt a tool in the first place.

First, the views of end users may differ dramatically from those of impacted communities. For example, in the child welfare context, an end user of a screening tool for child maltreatment allegations is the child welfare agency worker—not the family being scrutinized and at risk of being separated. In the criminal legal context, an end user of a risk assessment tool may be the judge, not the individual whose liberty, livelihood, and ability to care for their loved ones is at stake in the assessment tool's actual use. Or in the housing, employment, or credit context, the end user may be the housing provider, employer, or lender, but the impacted individuals are the people who may be denied a home, job, loan, or other basic economic opportunities. In addition to having questions about the tool's ability to make truly relevant individualized assessments, impacted individuals and communities have a direct, personal stake in how and when the tool is used.

Engaging impacted communities over the entire lifetime of an AI tool is not a matter only of fairness and efficacy, but also of realizing the promise of civic engagement in a democracy. When government agencies use algorithmic decision-making tools, the values and policies embedded in and enacted by those tools must be available to the public for review and debate. Especially when these tools are deployed in systems with preexisting racial, gender, disability, or age disparities, the failure to engage impacted communities, to disclose how a tool was created or how it works, and to provide meaningful opportunities to comment or object erodes public trust in AI systems' fairness and efficacy. To the extent these tools are used to make decisions that impact fundamental liberty interests or equal protection—such as child welfare agencies intruding into the parent-child relationship, [5] law enforcement detaining and incarcerating individuals, public housing authorities denying individuals access to housing/shelter, or administrative agencies denying individuals the ability to access employment opportunities—ensuring transparency and accountability by

[3] *Id.* at lines 476–78.
[4] *Id*. at lines 480–82.
[5] *See, e.g., Meyer v. Nebraska*, 262 U.S. 390, 399 (1923) (holding that the Due Process Clause of the Fourteenth Amendment protects "the right of the individual to . . . marry, establish a home and bring up children"); *Pierce v. Soc'y of Sisters*, 268 U.S. 510, 534-35 (1925) (recognizing the "liberty of parents and guardians to direct the upbringing and education of children under their control"); *Santosky v. Kramer*, 455 U.S. 745, 753 (1982) (acknowledging the "fundamental liberty interest of natural parents in the care, custody, and management of their child.").

meaningfully involving impacted communities in the development, adoption, and review of AI tools is of the utmost importance.

We recommend: (1) that NIST's final publication explicitly name impacted individuals and communities as key stakeholders in all aspects of the AI lifecycle, and (2) that NIST meaningfully involve impacted communities in its plans to develop a framework for trustworthy and responsible AI.

## II. Lack of Trust in AI is Linked to Lack of Trust in Institutions

**As a comment to:**

- Spec. Pub. 1270 lines 104–05: **"**Working with the AI community, NIST has identified the following technical characteristics needed to cultivate trust in AI systems"
- Spec. Pub. 1270 lines 196–98: "NIST research in AI continues along this path to focus on how to measure and enhance the trustworthiness of AI systems. Working with the AI community, NIST has identified the following technical characteristics needed to cultivate trust in AI systems"
- Spec. Pub. 1270 lines 208–09: "Specifically, how the presence of bias in automated systems can contribute to harmful outcomes and a public lack of trust."
- Spec. Pub. 1270 lines 269–73: "There have long been two common assumptions about the rise and use of automation: it could make life easier and also create conditions that reduce (or eliminate) biased *human* decision making and bring about a more equitable society. These two tenets have led to the deployment of automated and predictive tools within trusted institutions and high-stake settings."

Throughout the draft, managing AI bias is framed as a solution to remedying public mistrust. While this represents an important goal, NIST must acknowledge that the lack of public trust in AI systems is based on numerous examples, some cited in the draft, of untrustworthy AI deployed on unsuspecting or powerless populations. Framings such as "cultivate trust in AI systems"[6] put the onus on the public to increase their trust, rather than on institutions to prove they are worthy of public trust in their adoption and use of technology. Until recently, institutions have done little to engender public trust. Few companies or institutions have publicly released proof or evidence that good-faith efforts have successfully addressed public concerns.[7] In some instances, institutions have put out dubious or questionable efforts[8] in an attempt to create the appearance of

[6] NIST, *supra* note 1, at line 105.

[7] *See, e.g.*, Will Knight, *The Apple Card Didn't 'See' Gender—and That's the Problem*, Wired (Nov. 19, 2019), https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/ [https://perma.cc/QKH2-RB9G]; Shunyuan Zhang et al., *Can an AI Algorithm Mitigate Racial Economic Inequality? An Analysis in the Context of Airbnb* (Rotman Sch. of Mgmt., Working Paper No. 3770371, 2021), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3770371 [https://perma.cc/6P9T-JT9H].

[8] *See, e.g.*, Alex C. Engler, *Independent Auditors Are Struggling to Hold AI Companies Accountable*, Fast Company (Jan. 26, 2021), https://www.fastcompany.com/90597594/ai-algorithm-auditing-hirevue [https://perma.cc/BUC3-SZA7]; Letter from Rebecca T. Wallace, Senior Policy Counsel, ACLU of Colorado & Aaron Horowitz, Deputy Chief Analytics Officer, ACLU Nat'l to Hon. Brian J. Flynn, Chief Judge, 21st Jud. Dist. (Oct. 29, 2020), https://www.courts.state.co.us/userfiles/file/Court_Probation/21st_Judicial_District/2021-07%20Vacate%20Bond%20Guidelines%20AO%202018-01.pdf [https://perma.cc/MD82-E9DS].

trustworthiness, when that trustworthiness was not deserved—a practice commonly described as "ethics-washing."[9] It is thus concerning that the primary contributors to the publication are creators of AI systems and practices, and not reflective of a broader community of stakeholders. It is particularly concerning that the team of contributors does not include people and communities who would directly experience the effects of the use of AI tools. AI practitioners and researchers have not yet proven themselves worthy of public trust. Encouraging trust may itself be counterproductive, if it encourages trust in untrustworthy systems.

The lack of trust in AI is mentioned throughout the publication, as is building and gaining public trust. The authors make the claim that AI is deployed to "trusted institutions and high-stakes settings" as the result of a desire to make life easier and reduce or eliminate human bias. However, we often see the opposite in practice—AI is hailed as a techno-solutionist option to some of the *least* trusted institutions, because human bias is perceived as so problematic to the functioning of those institutions. Among institutions we know of that deploy AI/machine learning ("ML"), there is the criminal legal system, in which only 20% of Americans have a great deal or quite a lot of trust, large technology companies, in which only 29% of Americans have a great deal or quite a lot of trust, banks, in which only 33% of Americans have a great deal or quite a lot of trust, and police, in which 49% of Americans do not have a great deal or quite a lot of trust.[10] For people impacted by these institutions' existing biases—primarily communities of color—the level of trust is especially low.[11] It should come as no surprise that Black people are twice as likely to not trust their local police as the general population, and 1.6 times more likely to expect it to be common for innocent people to be convicted of a crime.[12] Federal policies that have codified and legalized biased treatment of marginalized communities have led to predictable disparities and a breakdown of trust in many public systems in the U.S. In those situations, adding AI to an already broken system can amplify its bias.[13]

A salient contributor to the lack of trust among institutions in the U.S. is those institutions' own lack of responsiveness to public opinion and critique. For example, much of the American banking and credit system relies on credit scores as reported by three private companies: Equifax, TransUnion, and Experian. These scores reflect racial and ethnic bias in historical access to credit and housing, making it more difficult for some to obtain housing or other lines of credit as a result

---

[9] *See, e.g.*, Yochai Benkler, *Don't let industry write the rules for AI*, Nature (May 1, 2019), https://www.nature.com/articles/d41586-019-01413-1 [https://perma.cc/9ZDX-PEFR].

[10] *See* Gallup, *Confidence in Institutions* (2021), https://news.gallup.com/poll/1597/confidence-institutions.aspx [https://perma.cc/MSZ6-XUQG].

[11] *See, e.g.*, Fed. Deposit Ins. Corp., *2017 FDIC National Survey of Unbanked and Underbanked Households* 20 (Oct. 2018), https://www.fdic.gov/householdsurvey/2017/2017report.pdf [https://perma.cc/X8UM-P3TU] (finding Black Americans are 5.3 times more likely to be unbanked).

[12] *See, e.g.*, Jamie Ballard, *Black Americans Less Likely to Feel Safe Speaking with Police*, YouGov (Nov. 18, 2018), https://today.yougov.com/topics/politics/articles-reports/2018/11/13/black-americans-police-safety-trust [https://perma.cc/N2XZ-NPXM].

[13] *See, e.g.*, Danielle Ensign et al., *Runaway Feedback Loops in Predictive Policing*, 81 Proc. Machine Learning Rsch. 1 (2018), http://proceedings.mlr.press/v81/ensign18a/ensign18a.pdf [https://perma.cc/Y8DH-6CJV]; Alex Albright, *If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions* (Sept. 3, 2019) (Ph.D. dissertation, Harvard University), https://thelittledataset.com/about_files/albright_judge_score.pdf [https://perma.cc/L96R-RTDX]; Robert Bartlett et al., *Consumer-Lending Discrimination in the FinTech Era* (Nat'l Bureau of Econ. Rsch., Working Paper No. 25943, June 2019), https://www.nber.org/system/files/working_papers/w25943/w25943.pdf [https://perma.cc/LWV8-XKDN].

of long-standing biased policies and lack of generational wealth.[14] Several relevant factors such as years of on-time utility and rent payments—arguably a significant contributor to habits in repaying debts—are not included in the traditional score despite public calls for this change.[15] In 2018, Experian created a service in response to this desire for scores to be more reflective of on-time payments of utility, phone bills, and even streaming subscriptions,[16] though opting into this service improves only the Experian credit score, leaving the Equifax and TransUnion scores unchanged. Additionally, credit scores, although touted as a means to model credit worthiness, actually can decrease when debts are paid off. As longstanding public critiques of the current systems continue to be ignored, lack of trust in institutions like credit-scoring remains reasonable, and that lack of trust will likely extend to future technologies. Until public opinion is respected and addressed, there will continue to be skepticism and mistrust of U.S. institutions.

AI is often deployed not in trusted institutions, but in institutions that have historically implemented (and continue to implement) systemically biased policies and actions, with consequent biased and harmful impacts. Algorithmic models are trained using data that emerges from these biased systems, and yet the algorithmic decisions are promoted in this context as a solution to this institutional untrustworthiness. The best course of action for NIST is to first seek to improve the institution, rather than push for additions of AI/ML to these settings. Whether AI is employed by these institutions or not, there is still a fundamental lack of trust that must be addressed.

## III. Transparency, Auditing, and Impact Assessments

**As a comment to:**

- Spec. Pub. 1270 lines 198–200: "NIST has identified the following technical characteristics needed to cultivate trust in AI systems: accuracy, explainability and interpretability, privacy, reliability, robustness, safety, and security (resilience)—and that harmful biases are mitigated."

While NIST's proposal focuses on technical characteristics, there is one non-technical characteristic that must accompany technical standards: transparency. Transparency is needed with respect to both the consumers, users, and/or targets of AI systems, as well as to the public in general. This first type of transparency comes in the form of explainability and interpretability,

---

[14] *See* Caroline Ratcliffe & Steven Brown, *Credit Scores Perpetuate Racial Disparities, Even in America's Most Prosperous Cities*, Urban Wire (Nov. 20, 2017), https://www.urban.org/urban-wire/credit-scores-perpetuate-racial-disparities-even-americas-most-prosperous-cities [https://perma.cc/ZTJ5-4JQ8]; *Who's Keeping Score? Holding Credit Bureaus Accountable and Repairing a Broken System: Hearing Before the H. Comm. on Fin. Servs.*, 116th Cong. (2019) (written testimony of Jennifer Brown, Assoc. Dir. of Econ. Pol'y, UnidosUS), https://www.congress.gov/116/meeting/house/108945/witnesses/HHRG-116-BA00-Wstate-BrownJ-20190226.pdf [https://perma.cc/8JSX-XYL2].

[15] *See* Natalie Campisi, *From Inherent Racial Bias to Incorrect Data—The Problems With Current Credit Scoring Models*, Forbes (Feb. 26, 2021), https://www.forbes.com/advisor/credit-cards/from-inherent-racial-bias-to-incorrect-data-the-problems-with-current-credit-scoring-models/ [https://perma.cc/WY34-THH6].

[16] *See* Brian Cassin, *Millions of American consumers will have the opportunity to instantly improve their credit score and get access to the credit they deserve*, Experian (Dec. 18, 2018), https://www.experian.com/blogs/news/2018/12/18/experian-boost/ [https://perma.cc/G9R7-5MPH]; *Experian Giving Consumers a Credit Boost for Their Love of Streaming*, Business Wire (July 27, 2020), https://www.businesswire.com/news/home/20200727005185/en/Experian-Giving-Consumers-a-Credit-Boost-for-Their-Love-of-Streaming [https://perma.cc/YS6G-ABLH].

which we understand is a topic about which NIST has already sought separate comments.[17] However, NIST's publication does not sufficiently address the second type of transparency. Transparency towards the public has the potential to minimize bias and cultivate trust in AI by opening the "black box" of algorithmic decision-making for stakeholders and giving outside parties the opportunity to continually evaluate AI systems for bias and discrimination.

First, AI must be evaluated through regularized audits and ongoing impact assessments. Currently too few individuals, organizations, and government regulators are able and willing to provide audits and impact assessments focused on bias. It is also unclear whether entities providing audits and impact assessments have meaningfully included input from impacted people and communities in designing and executing audits and impact assessments.[18] Audits and impact assessments should be conducted according to standards that set out necessary evaluation points, and at minimum should require: regular evaluation for discriminatory effects throughout the model's conception and development, and—if not terminated during development due to an unacceptable risk of bias or other reasons—in its implementation and use; proactive searches and adoption of less discriminatory alternatives; and assessments of whether data used in training technologies is representative and accurate, and that the technologies measure lawful and meaningful attributes and seek to predict valid target outcomes. While some entities are voluntarily evaluating their AI systems, there is not enough transparency in the documentation and publication of those audits and impact assessments. When an audit or impact assessment is conducted, it is important that information about the evaluation be publicly available, including information about the content and reasoning behind the evaluation, who is conducting the evaluation, and what their relationship is to the entity being evaluated. Audits and impact assessments should ideally only be conducted by independent third-party actors who do not have a stake in whether a system will ultimately be utilized or not, but in some instances, internal audits by neutral actors with no stake in whether a system is ultimately used may also be valuable. Additionally, the entity should be transparent about the scope of the audit or impact assessment.[19] Results should also be made public. Audits and impact assessments can be coupled with the technical counterfactual techniques and other technical bias-reducing measures NIST has already proposed to help mitigate bias.[20] Standards will also need to lay out what kind of information should be retained and documented about the technology, its development, and internal auditing sufficient to allow for third party auditing.

---

[17] *See* P. Jonathon Phillips et al., Nat'l Inst. of Standards & Tech., Draft NISTIR 8312, Four Principles of Explainable Artificial Intelligence (Aug. 2020), https://nvlpubs.nist.gov/nistpubs/ir/2020/NIST.IR.8312-draft.pdf [https://perma.cc/2JBE-RADY]; David A. Broniatowski, Nat'l Inst. of Standards & Tech., NISTIR 8367, Psychological Foundations of Explainability and Interpretability in Artificial Intelligence (Apr. 2021), https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8367.pdf [https://perma.cc/KL8N-VKCW].

[18] Dillon Reisman et al., *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability*, AI Now Inst. 18–20 (Apr. 18, 2018), https://ainowinstitute.org/aiareport2018.pdf [https://perma.cc/JQY8-VB4L] ("[Entities] should ensure that affected communities are able to suggest researchers that they feel represent their interests, and should work with researchers to ensure that these communities have a voice in formulating the questions that are asked and addressed by research and auditing.").

[19] Alex Engler, *Auditing Employment Algorithms for Discrimination*, Brookings Inst. (Mar. 12, 2021), https://www.brookings.edu/research/auditing-employment-algorithms-for-discrimination/ [https://perma.cc/JMT5-TUSJ] (describing the various forms and scopes of audits, including a recent audit of HireVue, Inc. which "did not independently analyze HireVue's data or directly evaluate its models").

[20] NIST, *supra* note 1, at lines 654–64.

Second, to the maximum extent possible, the public needs access to detailed information about entities' uses of AI, any audits and impact assessments of those models, and any relevant agency reviews. This is especially true in the fields of housing, employment, and credit ("HEC"), where AI decision-making can have lifechanging consequences for people. NIST can facilitate such transparency by, for example, producing research reports in partnership with directly impacted communities, civil rights organizations (including local direct service/organizing organizations and national organizations), consumer protection groups, non-profit research agencies, and HEC institutions using AI.[21] NIST can also push for the results of audits and impact assessments to be made public.

Additional important questions for which there are not currently standards include transparency around what information underlying an audit or impact assessment must be archived for regulatory review and how frequently audits and impact assessments should be conducted.

The problem of a lack of transparency thus cannot be solved with technical standards alone. It is crucial that NIST's proposed technical standards be coupled with comprehensive guidance, legislation, and regulation to ensure that AI systems mitigate bias and earn public trust. In particular, NIST should emphasize the need for meaningful inclusion of people and communities that will bear the effects of the deployment of AI in discussions about how audits and impact assessments should be structured and executed.

## IV. Algorithmic Responsibility and the Relationship with Policy

**As a comment to:**

- Spec. Pub. 1270 lines 428–430: "Decisions here include how to frame the problem, the purpose of the AI component, and the general notion that there is a problem requiring or benefitting from a technology solution"
- Spec. Pub. 1270 lines 512–515: "This stage of the AI lifecycle is where modeling, engineering and validation take place. The stakeholders in this stage tend to include software designers, engineers, and data scientists who carry out risk management techniques in the form of algorithmic auditing and enhanced metrics for validation and evaluation."

NIST's draft publication does an excellent job of highlighting the many different points in the algorithmic decision-making process at which bias can creep into models, from choosing which variables to include to identifying a reasonable outcome variable. All of these are critical points where harm can occur. However, especially in contexts where algorithmic decision-making is deployed for government operations, these are also *policy* decisions hidden within a process that remain obscure to policymakers. In the process of making algorithmic decisions, the people who are designing an algorithm often also make policy decisions. The decision-making around policy that occurs when designing algorithms often goes unexamined and consequently, the authority and expertise of creators of algorithms to set policy goes unexamined. For instance, for pretrial risk assessments, which are used in determining whether to hold a person in jail while they await trial, current algorithms primarily attempt to predict re-arrest or failure to appear in court. However, the only compelling interests that might justify pre-trial detention are preventing pretrial flight and

[21] *See, e.g.*, Nat'l Inst. of Standards & Tech., *NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software* (Dec. 19, 2019), https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software [https://perma.cc/M568-MB56].

violent crime.[22] These proxy variables are *not* the same,[23] and the decision of government actors to use them in the pretrial context should be subject to the same rule-making processes as if these factors were used outside the context of an algorithm.

Furthermore, in the same context, algorithm designers often choose (sometimes with criminal legal system actors such as pretrial services departments, police, and judges, but rarely with public defenders, defendants themselves, or representatives of heavily policed communities) the thresholds between what makes someone "high risk" and "low risk" based on the probabilities the models generate. But these designations are not reliable at an individual level.[24] Since *United States v. Salerno* establishes that liberty before trial is a fundamental right and states that "detention prior to trial or without trial is the carefully limited exception," policymakers should be setting these thresholds in line with the constitutional norm and with full knowledge of the public safety and pretrial liberty tradeoffs in play.[25] But that is rarely the case.[26] The Bureau of Prisons implicitly acknowledged algorithm thresholds as policy when they altered the threshold of what makes a person "low risk" just in time to avoid releasing vulnerable incarcerated people from prison who, if they'd been assessed by the same instrument in the year prior, would have been released.[27] If a policymaker can modify operating conditions of an algorithm *after* it is deployed, then decisions made by algorithm designers themselves are also policy choices.

On the other hand, policymakers have yet to do critical work required for tool-developers to effectively make any assurances that their work is minimizing harm.[28] In some areas, there is an attempt to address disparate impact through laws and regulations such as the Fair Housing Act, or agency regulations (e.g., Equal Employment Opportunity Commission, U.S. Department of Housing and Urban Development, Federal Trade Commission, or Consumer Financial Protection Bureau regulations). Yet in many other areas, no such laws exist, and no policymakers have put in place any guardrails to ensure the definition of harm is appropriate to the context. And then there are some laws that actually *prevent* and *inhibit* our ability to minimize harm from bias. Pre-processing or in-processing techniques in credit and lending scenarios are effectively prohibited by Regulation B of the Equal Credit Opportunity Act, which does not allow the use or collection

---

[22] *See* Brandon Buskey & Andrea Woods, *Making Sense of Pretrial Risk Assessments*, The Champion (June 2018), https://www.nacdl.org/Article/June2018-MakingSenseofPretrialRiskAsses [https://perma.cc/Q3S7-QCG4]; *see generally Stack v. Boyle*, 342 U.S. 1 (1951); *United States v. Salerno*, 481 U.S. 739 (1987).

[23] *See* Riccardo Fogliato et al., *On the Validity of Arrest as a Proxy for Offense: Race and the Likelihood of Arrest for Violent Crimes* (May 11, 2021) (preprint), https://arxiv.org/pdf/2105.04953.pdf [https://perma.cc/T79X-XASQ]; Lauryn P. Gouldin, *Defining Flight Risk*, 85 U. Chi. L. Rev. 677 (2018), https://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=6089&context=uclrev [https://perma.cc/NQD3-PEG8]; Ethan Corey & Puck Lo, *The 'Failure to Appear' Fallacy*, The Appeal (Jan. 9, 2019), https://theappeal.org/the-failure-to-appear-fallacy/ [https://perma.cc/G273-X8TG].

[24] *See e.g.*, Kristian Lum et al., *Closer than they appear: A Bayesian Perspective on Individual-Level Heterogeneity in Risk Assessment* (Feb. 1, 2021) (preprint), https://arxiv.org/pdf/2102.01135.pdf [https://perma.cc/8KYE-X2RA].

[25] 481 U.S. at 755.

[26] *See* Megan T. Stevenson & Sandra G. Mayson, *Pretrial Detention and the Value of Liberty* (Va. Pub. L. and Legal Theory Rsch. Paper No. 2021-14, Feb. 16, 2021), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3787018 [https://perma.cc/WRS7-LW3M].

[27] *See, e.g.*, Ian MacDougall, *Bill Barr Promised to Release Prisoners Threatened by Coronavirus—Even as the Feds Secretly Made it Harder for Them to Get Out*, ProPublica (May 26, 2020), https://www.propublica.org/article/bill-barr-promised-to-release-prisoners-threatened-by-coronavirus-even-as-the-feds-secretly-made-it-harder-for-them-to-get-out [https://perma.cc/FT8Y-E7SF].

[28] *See, e.g.*, Rashida Richardson, *Defining and Demystifying Automated Decision Systems*, 81 Md. L. Rev. ___ (forthcoming 2022), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3811708 [https://perma.cc/DKU6-RFBA].

of protected characteristics, such as race or gender, by financial institutions except for mortgage decisions.[29] The entire process of assuring least discriminatory alternatives is built off of imputing race and gender based on location,[30] but the ability to impute this information is itself the result of an extensive history of racial discrimination in housing—only managing bias in AI may not be able to fix this long-standing issue.[31] In the voting rights context, laws have been created that explicitly *will* prevent algorithms that could be used to manage bias and protect civil liberties. For example, in Georgia, the "exact match" law requires perfect name matching between government issued-IDs and the voter roll. The use of probabilistic record-linking thus no longer applies in this context even though it would assure the right to vote, especially for minority communities which are more likely to be purged by exact match laws.[32]

## V. AI Systems Are Often Implemented for Economic Reasons

One factor that can lead to bias is the economic incentive that accompanies the adoption of an AI system. For example, an agency that needs to disburse benefits to some subset of the public while vetting applicants might traditionally employ civil servants to review individual applications, discuss with the applicants, and determine the appropriate benefits. Not only are benefits themselves costly, but civil servant trainings, salaries, and support can be a significant cost. Economic pressure might incentivize agencies to adopt measures that can relieve some of the workload from their staff and result in a reduction of overall payout.

For example, Idaho's Medicaid program moved to an automated system to apportion benefits, presumably in part to relieve the burden on assessors of determining necessary amounts, enabling each assessor to process more applicants in the same amount of time.[33] In this case, little review was done on the effectiveness or quality of the automated system before deployment, and the warnings raised by what little review was done were ignored. Benefits were reduced for many applicants, and those reductions were disproportionate across different parts of the state.

Medicaid itself is distributed disproportionately across gender and other protected characteristics. In Idaho for example, 55% of beneficiaries are women[34] while women make up

---

[29] *See, e.g.*, Equal Credit Opportunity Act (Regulation B) Ethnicity and Race Information Collection, 12 C.F.R. § 1002 (2017), https://www.federalregister.gov/d/2017-20417 [https://perma.cc/VR8C-D2AH].

[30] *See e.g.*, Consumer Fin. Prot. Bureau, *Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity* (2014), https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf [https://perma.cc/Z3ZS-8Y9S].

[31] *See* Laura Blattner & Scott Nelson, *How Costly is Noise? Data and Disparities in Consumer Credit* (May 5, 2021) (preprint), https://arxiv.org/abs/2105.07554 [https://perma.cc/N3ZA-L57Y]; Rashida Richardson, *Racial Segregation and the Data-Driven Society: How Our Failure to Reckon with Root Causes Perpetuates Separate and Unequal Realities*, 36 Berkeley Tech. L.J. ___ (forthcoming 2022), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3850317 [https://perma.cc/5RBR-VZVK].

[32] *See e.g.*, Ted Enamorado, *Georgia's 'Exact Match' Law Could Potentially Harm Many Eligible Voters*, Wash. Post (Oct. 20, 2018), https://www.washingtonpost.com/news/monkey-cage/wp/2018/10/20/georgias-exact-match-law-could-disenfranchise-3031802-eligible-voters-my-research-finds/ [https://perma.cc/475V-GTBC].

[33] Jay Stanley, *Pitfalls of Artificial Intelligence Decisionmaking Highlighted in Idaho Case*, ACLU (June 2, 2017), https://www.aclu.org/blog/privacy-technology/pitfalls-artificial-intelligence-decisionmaking-highlighted-idaho-aclu-case [https://perma.cc/J4J3-GUNS].

[34] Kaiser Fam. Found., *Medicaid Enrollment by Gender*, https://www.kff.org/medicaid/state-indicator/medicaid-enrollment-by-gender/?currentTimeframe=0&selectedRows=%7B%22states%22:%7B%22idaho%22:%7B%7D%7D%7D&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D [https://perma.cc/TF3E-PX8T] (last visited Sept. 9, 2021).

only 50% of the state's population.[35] Even if benefits had been reduced proportionately by gender *within* Medicaid in Idaho, those losses would have fallen disproportionately on women simply because women are overrepresented in the population who are subject to this algorithmic judgment. When automated systems are adopted in the face of economic pressures, bias and equity should be evaluated not only within a particular domain, but also in the surrounding social context. The analysis must ask not only, "Is this an equitable mechanism for gaining cost savings in this application?" but also, "Are there other ways we could achieve the same cost savings without the proposed system? Would the other ways be more equitable?"

Ironically, the Medicaid Act already requires explanations for any benefit reduction, but those explanations were lacking in Idaho because administrative staff were unable to interpret their own decision-making system. It required a lawsuit by the ACLU of Idaho to bring the system into compliance with the transparency requirement, and when the underlying decision-making system was analyzed, its internal processes were found to be scientifically unjustified.[36] Part of being able to address bias in an automated decision-making system is just this kind of transparency and explainability, as well as a responsive, well-staffed, and impartial process for appealing AI-assisted decisions. All of these critical safeguards themselves come with system design, labor, and maintenance costs, not to mention the additional costs of compensating affected parties when an automated decision is found to have been misjudged.

To the extent that adoption of any automated decision-making system is based on economic pressures, NIST must make clear that the costs of appropriate oversight, safeguards against bias, and compensation for misjudgment must be included in the economic analysis of the algorithmic decision-making system as a whole. When estimates or measurements of these costs overwhelm the intended economic benefits, it should be clear that the proposed system should be avoided or decommissioned.

## VI. The AI Lifecycle: NIST Must Model Scenarios Where AI Should Not Be Used

Given the many sources of bias and their potential real-world impact, NIST's framework must give far greater attention to modeling scenarios in which AI should not be used at all—or, if already deployed, should be removed from use. The working assumption of the proposal is that biases in AI can be identified and adequately mitigated to allow for the responsible deployment of AI systems. But in many cases that assumption may be unjustifiable. NIST's AI lifecycle must explicitly model decision-making that culminates in a determination that the deployment of particular AI tools would be harmful and/or improper.

NIST's proposal acknowledges that certain AI tools should not be developed or deployed at all, but it does so only in passing: "[I]t may become apparent that algorithms are biased or will contribute to disparate impacts if deployed. In such cases the technology can be taken out of production."[37] The proposal goes on to highlight exactly why clear, well-developed guidance on *this* type of decision-making—a decision to halt development or deployment of AI—is urgently needed from NIST. "[T]his kind of awareness and remedy is likely to take place only in certain settings or industries, with well-defined procedures and clear lines of accountability.

[35] U.S. Census Bureau, *QuickFacts Idaho* (2019), https://www.census.gov/quickfacts/ID [https://perma.cc/N3K9-UG4S].
[36] *See K.W. ex rel. D.W. v. Armstrong*, 789 F.3d 962 (9th Cir. 2015).
[37] NIST, *supra* note 1, at lines 535–37.

Unfortunately, not all tools are deployed in such settings—and capturing the wide array of use cases and scenarios is particularly difficult."[38]

The key diagram that NIST has used to illustrate the AI lifecycle (Figure 1) does not model decision-making—or even acknowledge the possibility—that the development or use of AI should in certain cases be terminated.[39] Instead, the figure contemplates a process of gradual refinement based on stakeholder input, risk management, and standards development with the goal of reducing bias—a process that, as represented in Figure 1, invariably culminates in deployment. This vision of the AI lifecycle is biased in favor of AI. Instead, NIST should revise its representation of the AI lifecycle to reflect the multiple points at which the developers of algorithms must consider whether the danger of bias in a given application is intolerable, and therefore whether the AI tool in question should not be used. In particular, the AI lifecycle must model two critical scenarios: (1) where the development of an algorithm is terminated *before* deployment, after assessing potential biases and the resulting impacts; and (2) where the use of an algorithm is terminated *after* deployment, because real-world use and ongoing auditing has revealed unacceptable biases and impacts.

Beyond this key depiction of the AI lifecycle, NIST must provide researchers and practitioners with a more developed framework for assessing whether the existence or risk of bias is intolerable in a given application. In developing this framework, NIST should intentionally seek out meaningful input from the communities that will be affected by a given algorithmic deployment. That input should include the nature and magnitude of the risks as they are experienced by individuals and whether the potential harms associated with the algorithm outweigh or otherwise complicate assessments of the potential benefits. NIST's proposal explains that the goal of its framework is "not 'zero risk,' but to manage and reduce bias in a way that contributes to more equitable outcomes that engender public trust."[40] While this goal is understandable, the framework does not provide concrete guidance or specific examples where the risk of bias is so severe or so unavoidable that the use of an AI tool is unacceptable. Nor does it include specific examples of who should be involved in making such a determination. Those developing and deploying AI systems—including both private companies and government agencies—badly need guidance in this area.

Many sources of bias may be impossible to remedy in practice, as NIST's proposal suggests. For example, "[o]ne challenge rests on the reality that decisions about which data to use . . . are often made based on what is available or accessible, rather than what might be most suitable—but difficult or impossible to utilize."[41] Moreover, "[e]ven if datasets are reflective of the real world, they may still exhibit entrenched historical and societal biases, or improperly utilize protected attributes."[42] For example, zip code data is strongly correlated with race based on a long history of residential segregation in the United States. Zip code information can therefore often function as a proxy for race, resulting in discriminatory decision-making.[43] Depending on what datasets are

[38] *Id.* at lines 537–40.
[39] *Id.* at line 415 (fig. 1).
[40] *Id.* at lines 352–53.
[41] *Id.* at lines 289–91.
[42] *Id.* at lines 301–02.
[43] Alexandra George, *Thwarting Bias in AI Systems*, Carnegie Mellon Univ. (Feb. 4, 2019), https://www.ece.cmu.edu/news-and-events/story/2019/02/thwarting-bias-in-ai-systems.html [https://perma.cc/RNC4-WX8W]; Jeff Larson et al., *How We Examined Racial Discrimination in Auto Insurance Prices*, ProPublica (Apr. 5, 2017), https://www.propublica.org/article/minority-neighborhoods-higher-car-insurance-premiums-methodology [https://perma.cc/UBV8-K2SG].

available, and what biases are embedded in them, it may not always be possible to "manage and reduce bias" to create algorithms that are genuinely deserving of public trust. NIST's framework should explicitly say so.

There are already real-world deployments of algorithms that are so plagued by bias they should either be removed from use or urgently reexamined.

The use of algorithms to generate individualized risk assessment scores for bail determinations or recidivism predictions for probation or parole decisions are glaring examples. These algorithms are used to make fundamental decisions about a person's liberty—including decisions about whether a person will be locked up prior to trial or released on bail, and whether a person will be subjected to ongoing supervision by law enforcement officers—with immense consequences for their ability to care for family members, remain employed, maintain stable housing, and receive adequate medical care. Yet the data used to train these algorithms reflects many of the flaws and biases identified by NIST in its proposal. This data may be incomplete because of poor record-keeping, inaccessibility, or incompatibility across various jurisdictions. And even if the data is relatively comprehensive, it reflects the operation of a criminal legal system that is structurally biased against communities of color, not only historically but to this day.[44] Such data primarily captures the behavior and decisions of police officers and prosecutors, acting on longstanding social biases, rather than the individuals or groups the data is claiming to describe. Given the biases that plague the underlying data and the life-altering consequences for people subject to these algorithms, there is an urgent need for NIST to model decisions by practitioners and policymakers to take algorithms offline.[45]

The same is true for algorithms used to make or guide child welfare decisions. AI is already being deployed in the child welfare context at various stages of the investigation process with little advance warning or disclosure to the public, let alone meaningful involvement of impacted families in evaluating utility or relevance to the actual needs of families facing child welfare involvement. Moreover, the child welfare system has been driven by expressly discriminatory and assimilationist policies throughout its existence.[46] As a result, women, Black and Indigenous families, and families in poverty have been and continue to be disproportionately surveilled and separated by the state, even where such policies have been disavowed, creating skewed datasets.[47]

---

[44] *See, e.g.*, Sandra Gabriel Mayson, *Bias In, Bias Out*, 128 Yale L. J. 2218 (2019), https://www.yalelawjournal.org/pdf/Mayson_p5g2tz2m.pdf [https://perma.cc/JBE5-9Y94]; John Logan Koepke & David G. Robinson, *Danger Ahead: Risk Assessment and the Future of Bail Reform*, 93 Wash. L. Rev. 1725 (2018), https://ssrn.com/abstract=3041622 [https://perma.cc/D85R-Q5LJ]; Laurel Eckhouse et al., *Layers of Bias: A Unified Approach for Understanding Problems With Risk Assessment*, 46 Crim. Just. & Behav. 185–209 (2019), https://doi.org/10.1177/0093854818811379 [https://perma.cc/3ERC-43NS]; Michelle Bao et al., *It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks* (June 10, 2021) (preprint), https://arxiv.org/abs/2106.05498 [https://perma.cc/9CWG-EFVN].

[45] *See* Andrea Woods & Portia Allen-Kyle, *A New Vision for Pretrial Justice in the United States*, ACLU (Mar. 2019), https://www.aclu.org/sites/default/files/field_document/aclu_pretrial_reform_toplines_positions_report.pdf [https://perma.cc/4GGK-D3UZ].

[46] *See generally* Dorothy Roberts, Shattered Bonds: The Color of Child Welfare (2001); Mical Raz, Abusive Policies: How the American Child Welfare System Lost its Way (2020).

[47] *See* Children's Bureau, *Child Welfare Practice to Address Racial Disproportionality and Disparity* 2–3 (2021), https://www.childwelfare.gov/pubPDFs/racial_disproportionality.pdf [https://perma.cc/PC79-FEGL]; *see also* Jerry Milner & David Kelly, *It's Time to Stop Confusing Poverty With Neglect*, The Imprint (Jan. 17, 2020), https://imprintnews.org/child-welfare-2/time-for-child-welfare-system-to-stop-confusing-poverty-with-neglect/40222 [https://perma.cc/DD42-ESCJ].

In the words of NIST's publication, the use of AI in this arena is precisely one that presents the "obvious risk" of "build[ing] algorithmic-based decision tools for settings already known to be discriminatory."[48]

Furthermore, because child maltreatment is frequently an unobservable or hard-to-observe event, prediction tools use proxies with poor construct validity,[49] such as the likelihood of a child's removal from their home within the next two years.[50] However, just as in the criminal legal context, proxies like the rate of child removal reflect the decisions and conduct of child welfare workers and government agencies, not parental or child behavior. The risk of harm by using discriminatory, inaccurate, or unreliable tools in child welfare and family regulation is great: while much attention has recently been paid to the long-lasting trauma that family separation causes to both a child and their parent,[51] even less drastic state intrusion into the parent-child relationship has been shown to cause emotional injury.[52] Meanwhile, in jurisdictions where AI tools are used in child welfare decision-making, families are not told that a score has been assigned to them, let alone what that score was or how an agency worker factored it into their decision-making. Thus, before AI is used in this context, the threshold question of whether AI should even be used to help the state decide which families to regulate and potentially separate should be extensively addressed.

Finally, NIST's framework must make clear that the decision to tolerate bias will often depend on the specific use case and the potential consequences for those affected—and thus is almost always a normative decision. Unless these normative judgments are made explicit as part of the design and auditing process, technical efforts to reduce and manage bias may simply serve to legitimize systems that continue to have profoundly inequitable and unfair results.

---

[48] NIST, *supra* note 1, at lines 447–50.

[49] *See* Abigail Z. Jacobs et al., *The Meaning and Measurement of Bias: Lessons from Natural Language Processing*, 2020 Proc. Assoc. for Computing Machinery Conf. on Fairness, Accountability & Transparency (Jan. 27, 2020), https://dl.acm.org/doi/abs/10.1145/3351095.3375671 [https://perma.cc/HMH3-TKZA].

[50] *See, e.g.*, Rhema Vaithianathan et al., Allegheny Family Screening Tool: Methodology, Version 2, at 2–3 (2019), https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/Methodology-V2-from-16-ACDHS-26_PredictiveRisk_Package_050119_FINAL-7.pdf [https://perma.cc/KLJ9-AHAB]; Oregon Dep't of Hum. Servs., Oregon DHS Safety at Screening Tool – Development and Execution at 3 (2019), https://www.oregon.gov/DHS/ORRAI/Documents/Safety%20at%20Screening%20Tool%20Development%20and%20Execution%20Report.pdf [https://perma.cc/X4ZT-YR56].

[51] *See* Kimberly Howard et al., *Early Mother-Child Separation, Parenting, and Child Well-Being in Early Head Start Families*, 13 Attachment & Hum. Dev. 5 (2011), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3115616/ [https://perma.cc/4WBJ-9WJM] (finding that mother-child separations of a week or longer within the child's first two years of life, for any reason, were associated with increased child negativity and aggression years later); Allison Eck, *Psychological Damage Inflicted By Parent-Child Separation is Deep, Long-Lasting*, PBS (June 20, 2018), https://www.pbs.org/wgbh/nova/article/psychological-damage-inflicted-by-parent-child-separation-is-deep-long-lasting/ [https://perma.cc/NXD9-BWDQ] (interviewing researchers who explain that forced family separation causes trauma with long-term consequences); Melissa de Witte, *Separation from Parents Removes Children's Most Important Protection and Generates a New Trauma, Stanford Scholar Says*, Stanford News (June 26, 2018), https://news.stanford.edu/2018/06/26/psychological-impact-early-life-stress-parental-separation/ [https://perma.cc/Q67S-Z2G4].

[52] Indeed, even families who are not involved in the system are harmed by the fear of future involvement. *See* Kenya Franklin & Careena Farmer, *New Research: How Fear of CPS Harms Families*, Rise Mag. (Jan. 22, 2020), https://www.risemagazine.org/2020/01/how-fear-of-cps-harms-families/ [https://perma.cc/YFU2-6Q8Z]; Kelley Fong, *Concealment and Constraint: Child Protective Services Fears and Poor Mothers' Institutional Engagement*, 97 Social Forces 1785 (2019), https://academic.oup.com/sf/article-abstract/97/4/1785/5113162 [https://perma.cc/4DRE-EPXN].

## VII. Data Privacy, Collection Methods, and Data Accuracy as Contributors to Algorithmic Bias

The publication acknowledges that the use of inherently biased datasets can contribute to harmful biases in AI systems early in the AI lifecycle. But the document underemphasizes (1) best practices for data collection that will be used to develop AI tools and how they differ from other data collection practices, (2) public hesitancy to contribute their data for fear of misuse, exploitation, and/or lack of adequate data privacy protections, and (3) research methods to mitigate bias in datasets early enough in the lifecycle to decide whether it is possible to build the AI tool given the available data.

The publication emphasizes the issue of bias in AI models while underemphasizing the role of poor data collection in creating models with harmful biases. Data collected for one purpose is often repurposed for use in AI systems. This data may have been collected in a way that optimizes its function for administrative or documentation purposes but may make it ill-fitting for the development of AI models. Alternatively, data might be used in a predictive model to make a determination far-flung from what the data represents in the real world, leading to bias and inaccuracy. An example of the latter occurred when an algorithm designed to determine how many vaccine doses to send to different locations in California was found to exclude millions of vulnerable people.[53] The state tried to ensure greater equity by focusing on zip codes that it found suffered worse during the pandemic by increasing the vaccine allotment to these areas. Many government agencies collect detailed demographic data on the people who live in each zip code, so it was assumed that zip codes could help meaningfully target vaccine allocation to hard-hit populations. The resulting issues with the algorithm, however, are rooted in what zip codes represent in the real world—geographic areas optimized for the sorting and delivering of mail. While these can be repurposed to glean population data, using zip codes to target vaccine distribution left behind residents of some smaller neighborhoods located within the boundaries of a particular zip code. In short, this data lacked the necessary intra-zip code granularity to achieve its intended goal.

A similar example of how poor and inaccurate data collection methods can introduce inaccuracy and bias into AI tools can be seen in predictive models used in medicine. Medical billing codes "are recorded by physicians when diagnosing a patient with a condition, and are used to ensure proper billing and insurance reimbursement."[54] These codes in electronic health records ("EHRs") are primarily meant to be used for clinical billing purposes and rely on human inputs which can reflect inaccuracies. This is due to the addition of codes to an EHR to issue and receive insurance reimbursement for a test to screen a patient for a disease; thus, relying solely on the codes in a patient's EHR is likely to introduce false positives. Moreover, some codes appear more often in the data simply because they are more easily reimbursed to the clinical staff—even if they don't represent a patient's accurate condition. In this example, humans have a direct role in generating incorrect data that will later be used to model patients' diagnostic risks. While the data achieves the administrative function of getting the medical clinic funds and providing patients with a health record of tests they received,

---

[53] *See* Brian Krans, *How Flaws in California's Vaccine System Left Some Oaklanders Behind*, The Oaklandside, (May 18, 2021), https://oaklandside.org/2021/05/18/how-flaws-in-californias-vaccine-system-left-some-oaklanders-behind/ [https://perma.cc/B6ZC-EGXP].

[54] *See* Brett Beaulieu-Jones, *Machine Learning for Structured Clinical Data*, *in* Advances in Biomedical Informatics 35–51 (Holmes & Jain eds., 2018), https://arxiv.org/abs/1707.06997 [https://perma.cc/NZ5G-672T].

it is not appropriate for the creation of AI/ML tools meant to, for example, model patients' health risks. Additionally, EHRs may contain duplicated patient data. While this may not impede clinical care, as all data belonging to one patient can be access and reviewed by a provider, data duplication can be problematic in algorithmic design, because an algorithm tested on the same data on which it was trained will suffer from overfitting.[55]

Another contributor to inaccuracy in EHR data, especially in prescription drug monitoring programs ("PDMPs"), may arise because medications for pets are listed under their owners' names.[56] In PDMPs, "a higher risk score correlates with increased probability that the prescribing or dispensing of a particular drug to a particular patient will result in negative consequence."[57] While the details of how PDMP risk score algorithms calculate this risk remain unknown because these tools are proprietary, it is known that the score incorporates a person's prescription history, physicians they've visited to detect "doctor shopping," and their current list of medications, among other factors. In addition to the possibility of pets' medications being misattributed to their owner, contributing to algorithmic performance issues, its design also appears to reinforce bias. In addition to the relevant medical information included in the algorithm, it also includes information on a patient's name, age, gender, address, prescription history, method of payment, distance travelled to provider and dispenser, drug-related arrests and convictions, child welfare cases, criminal cases, drug court case information, and more.[58] Because of historical biases, ethnic and racial minorities and socio-economically disadvantaged individuals may then see their patient risk scores artificially inflated as a result of these proxies.

Studies have found that individuals from vulnerable populations, including people in poverty, those with mental health disabilities, and immigrants, are more likely to visit multiple institutions and/or health care systems to receive clinical care.[59] Some PDMP algorithms may flag this as "doctor shopping" when it actually arises from more complex social circumstances. These same vulnerable patients may feature more missing data from their EHR as a result of changing health care institutions; they may be more likely to have been seen in teaching clinics, where data input and clinical diagnosis may be less accurate or systematically different; and they may be part of a sample too small for predictive clinical algorithms to generate accurate predictions. Disparate impact may also be experienced by patients with chronic pain and other stigmatized conditions—for example, PDMPs often reflect higher scores for patients with cancer.[60] This is an example of

---

[55] *See, e.g.*, Will Douglas Heaven, *Hundreds of AI Tools Have Been Built to Catch Covid. None of Them Helped.*, MIT Tech. Rev. (July 30, 2021), https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/ [https://perma.cc/DL8E-4B5G].

[56] *See* Maia Szalavitz, *The Pain Was Unbearable. So Why Did Doctors Turn Her Away?* Wired (Aug. 11, 2021), https://www.wired.com/story/opioid-drug-addiction-algorithm-chronic-pain/ [https://perma.cc/SGU5-DQQE].

[57] Jennifer Oliva, *Dosing Discrimination: Regulating PDMP Risk Scores*, 110 Cal. L. Rev. ___ (forthcoming 2022), http://dx.doi.org/10.2139/ssrn.3768774 [https://perma.cc/8FKV-YCBP]; *see also* Brief for Am. Civil Liberties Union et al. as Amici Curiae Supporting Respondent, U.S. Dep't of Just. v. Ricco Jonas, No. 19-1243 (1st Cir. argued Oct. 10, 2019) (describing PDMPs).

[58] Jennifer Oliva, *Dosing Discrimination: Regulating PDMP Risk Scores*, 110 Cal. L. Rev. ___ (forthcoming 2022), http://dx.doi.org/10.2139/ssrn.3768774 [https://perma.cc/8FKV-YCBP].

[59] *See* Milena A. Gianfrancesco et al., *Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data*, 178 JAMA Intern. Med. 1544 (2018), https://doi.org/10.1001/jamainternmed.2018.3763 [https://perma.cc/2DPH-UCX4].

[60] *See* Maia Szalavitz, *The Pain Was Unbearable. So Why Did Doctors Turn Her Away?* Wired (Aug. 11, 2021), https://www.wired.com/story/opioid-drug-addiction-algorithm-chronic-pain/ [https://perma.cc/SGU5-DQQE].

an algorithmic tool that uses data, even when accurate and not plagued by misattributions, to cause harm and suffering to the most vulnerable populations.

Bias in AI tools can be driven by data inaccuracy or data collection methods that do not optimize the data for use in AI models. NIST should encourage AI practitioners and government agencies to work to detect these factors and their effects on data accuracy in order to determine whether the data available can be used to create an accurate model free of harmful biases.

Another significant contributor to mistrust in AI tools may be driven by public mistrust in the collection of individuals' data. Past instances of poor data privacy protections, misuse of data, and outright mistreatment of members of some communities by research institutions, leading to their underrepresentation in future data, must be acknowledged as contributors to lack of public trust.

Examples of misuse of public data abound—these can range from instances in which private companies have profited off of public data without user consent, knowledge, or compensation, to instances where the sharing of people's identifying data put them at risk of harm. An example of the former is facial recognition databases that have been generated from images posted online without the consent of the posters or those pictured. In their effort to create a more diverse training dataset, IBM recently released a dataset with photos taken from the photo hosting site Flickr, for wider use by researchers.[61] For some, the photos in the dataset were annotated with information about the person pictured, including their ethnicity.[62]

This step of human labeling of training data is not itself without potential for bias. Some researchers note that the methods used in human annotation should be considered "a core aspect of the research process, with as much attention, care, and concern placed on the annotation process as is currently placed on performance-based metrics like F1 scores," since there can be variability in qualifications, training, and decisions made by the human labelers.[63] In attempts to minimize the cost of generating datasets, data laborers may receive minimal (if any) training in ethics when conducting annotation work, nor training on potential harmful social implications of publishing sensitive data.[64] Data labeling is rife with biases; for example, Facebook has flagged non-sexual content depicting same-sex couples as sexually "explicit," though equivalent content involving heterosexual couples is not flagged.[65]

---

[61] *See* Olivia Solon, *Facial Recognition's 'Dirty Little Secret': Millions of Online Photos Scraped Without Consent*, NBC News (Mar. 17, 2019), https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921 [https://perma.cc/5AA9-DBTZ]; John R. Smith, *IBM Research Releases 'Diversity in Faces' Dataset to Advance Study of Fairness in Facial Recognition Systems*, IBM Rsch. Blog (Jan. 29, 2019), https://www.ibm.com/blogs/research/2019/01/diversity-in-faces/ [https://perma.cc/UM6P-RC6V].

[62] *See* Shannon Liao, *IBM Didn't Inform People when it Used their Flickr Photos for Facial Recognition Training*, The Verge (Mar. 12, 2019), https://www.theverge.com/2019/3/12/18262646/ibm-didnt-inform-people-when-it-used-their-flickr-photos-for-facial-recognition-training [https://perma.cc/WHQ2-PS4P].

[63] *See* R. Stuart Geiger et al., *Garbage In, Garbage Out? Do Machine Learning Application Papers in Social Computing Report where Human-Labeled Training Data Comes From?* 2020 Proc. ACM Conf. on Fairness, Accountability & Transparency (2020), https://arxiv.org/abs/1912.08320 [https://perma.cc/J7XN-BLYS].

[64] *See* Morgan Klaus Scheuerman et al., *Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development*, 5 Proc. ACM on Hum.-Computer Interaction ___ (forthcoming Oct. 2021), https://arxiv.org/abs/2108.04308 [https://perma.cc/BS8H-CB5T].

[65] *See* Sera Golding-Young, *Facebook Blocked My Ad, Mislabeled it "Sexually Explicit*," ACLU of N. Cal. (Sept. 23, 2020), https://www.aclunc.org/blog/facebook-blocked-my-ad-mislabeled-it-sexually-explicit [https://perma.cc/CFX7-PVEF].

People's hesitancy to volunteer their data for the development of AI tools may be a direct result of high-profile instances in which private companies have collected data without user consent, knowledge, or compensation. An especially egregious and unethical example occurred with the revelation that Google, in an attempt to collect more data samples from Black participants in developing its facial recognition tool, allegedly scanned homeless Black people's faces.[66] Contractors collecting this data offered participants $5 gift cards and were allegedly encouraged "to approach homeless people, who [Google] expected to be most responsive to the gift cards and least likely to object or ask questions about the terms of data collection."[67] In some cases, contractors actively lied to participants, telling them that the phone scanning their faces was not recording them.

Data sharing within federal institutions has, at time, led to negative outcomes for vulnerable communities, as in 2018 when a data-sharing agreement between HHS and ICE actually led fewer potential sponsors for unaccompanied minors to identify themselves for risk of deportation.[68] Several research studies have found that facial recognition systems do not perform as well on Black people. Biased facial recognition has even led to wrongful arrests as algorithms have misidentified suspects.[69] In this case, a private company may have been attempting to address algorithmic bias, but did so in an unethical way that may have caused further public distrust in this AI tool.

Unethical data collection and research practices, whether by private companies or government institutions, can have a chilling effect on research participation among some populations. This in turn can lead to the collection of more homogenous data and AI tools that do not perform as well in some populations. Disturbingly, even when data is obtained unethically and the datasets are retracted, participants' data can still be widely used without their consent or knowledge.[70] Moreover, while research participants may agree to the use of their data given current technological tools, technological advances can change over time, raising new ethical concerns about the use of data in a way not initially intended or conceived of by research participants. And while participants may be comfortable sharing their data for research, the use of the same data for production may raise ethical concerns.

Considering the serious issues and questionable practices currently plaguing data stewardship, it is understandable why some, especially those in vulnerable communities, may not want their data collected in service of AI tools, especially those like facial or voice recognition that are disproportionately used for surveillance of marginalized communities. AI is a powerful

---

[66] *See* Jack Nicas, *Atlanta Asks Google Whether it Targeted Black Homeless People*, N.Y. Times, (Oct. 4, 2019), https://www.nytimes.com/2019/10/04/technology/google-facial-recognition-atlanta-homeless.html [https://perma.cc/7RKS-Q5W2].

[67] *See* Sidney Fussell, *How an Attempt at Correcting Bias in Tech Goes Wrong*, The Atlantic (Oct. 9, 2019), https://www.theatlantic.com/technology/archive/2019/10/google-allegedly-used-homeless-train-pixel-phone/599668/ [https://perma.cc/RL3J-N6VA].

[68] *See* Jonathan Blitzer, *To Free Detained Children, Immigrant Families Are Forced to Risk Everything*, The New Yorker (Oct. 16, 2018), https://www.newyorker.com/news/dispatch/to-free-detained-children-immigrant-families-are-forced-to-risk-everything [https://perma.cc/FJ4G-QPNV].

[69] Kashmir Hill, *Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match*, N.Y. Times (Jan. 6, 2021), https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html [https://perma.cc/HN7Q-XKPU].

[70] *See* Kenny Peng et al., *Mitigating Dataset Harms Requires Stewardship: Lessons from 1000 Papers* (2021) (preprint), https://arxiv.org/abs/2108.02922 [https://perma.cc/8HFM-XDRS].

tool but it requires large amounts of data that are representative of the population on which the AI tool will be applied. In pursuit of this data, AI practitioners have at times disregarded ethics considerations. In order to cultivate trust in AI tools, NIST must encourage practitioners to incorporate principles of ethics into the collection and protection of data that is used in the development of AI tools.

Given the possibility of inaccurate data and selection bias, methods[71] to counter sources of bias in datasets must be employed rigorously early in the AI lifecycle. Otherwise, the compounding effect of bad data may create situations in which biased AI tools encourage the perpetuations of biases in the real world. This effect is well documented in the finance world, where the history of racist policies and practices in, for example, the housing market, coupled with lack of regulation in banking, have impacted consumer behavior, leading to lower quality data for historically excluded groups like low-income people and minorities. A recent study—the largest ever on real-world mortgage data—found that algorithmic decisions around mortgage approval using credit scores were based on less data on the credit histories of minorities, and thus lacked precision disproportionately for members of minority groups.[72] In this case, a model with different levels of precision arose not from technical issues but from sociological factors. Technical standards alone cannot remedy the compounded effects of bad policies and practices here, as a lack of precision leads to fewer loan approvals for these groups, which then contributes to the paucity of accurate data about them—a vicious cycle.

Another example of the compounded effects of bad policies and practices with biased AI is evident in policing technologies. These technologies may reflect the problems with AI that likely cannot be remedied by technical standards. Many law enforcement agencies use predictive policing systems that are built on data known to be collected during periods of flawed, racially biased, and unlawful practices and policies—a practice referred to as "dirty policing."[73] Information generated from the use of these biased systems often influences the data that is input into them. If unchecked, a runaway feedback loop can result in which the same neighborhoods that have been heavily policed in the past are recommended for still heavier policing, regardless of actual reported crime.[74] These feedback loops not only perpetuate harm by seeming to validate entrenched policing policies, but they may also decrease the likelihood that other types of crime be curtailed. For example, many policing technologies focus on rare violent crimes, with no ability to predict more prevalent white-collar crimes.[75]

---

[71] *See* Michael Feldman et al., *Certifying and removing disparate impact*, Proc. of 21st ACM SIGKDD Conf. on Knowledge Discovery & Data Mining, (2015), https://arxiv.org/abs/1412.3756 [https://perma.cc/QQH3-VW7K].

[72] *See* Will Douglas Heaven, *Bias Isn't the Only Problem with credit scores—and no, AI can't help*, MIT Tech. Rev. (June 17, 2021), https://www.technologyreview.com/2021/06/17/1026519/racial-bias-noisy-data-credit-scores-mortgage-loans-fairness-machine-learning/ [https://perma.cc/3UQX-A9RK]; Laura Blattner & Scott Nelson, *How Costly is Noise? Data and Disparities in Consumer Credit* (May 5, 2021) (preprint), https://arxiv.org/abs/2105.07554 [https://perma.cc/9R8U-2TSD ].

[73] *See* Rashida Richardson et al., *Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice*, 94 N.Y.U. L. Rev. 15 (2019), https://www.nyulawreview.org/wp-content/uploads/2019/04/NYULawReview-94-Richardson_etal-FIN.pdf [https://perma.cc/78NL-AFGK].

[74] *See* Danielle Ensign et al., *Runaway Feedback Loops in Predictive Policing*, 81st Proc. of Machine Learning Rsch. Conf. on Fairness, Accountability & Transparency 1 (2018), http://proceedings.mlr.press/v81/ensign18a/ensign18a.pdf [https://perma.cc/Y8DH-6CJV].

[75] *See* Gerald Cliff & April Wall-Parker, *Statistical Analysis of White-Collar Crime*, Oxford Rsch. Encyclopedia of Criminology & Crim. Just. (Apr. 26, 2017), https://doi.org/10.1093/acrefore/9780190264079.013.267 [https://perma.cc/P3HJ-VE6Y].

"Automating all aspects of police information gathering, management and analysis has the risk of extending [a] techno-deterministic logic that posits machine-created information as more objective than human-generated information."[76] Moreover, enforcement based on this technology can appear erroneous at the individual level, when, for example, an individual who had never committed a crime was placed on a list of those likely to commit a crime, largely because of their geographic location.[77] Instances like this call into question the financial burden, ethics, and logic of attempting to make individual behavioral forecasts using population-level data. In the development of AI tools, especially those that purport to predict individual behavior, NIST must ensure that adequate steps have been taken to address bias in the data being used and that any feedback loops perpetuated by AI systems are mitigated. Given the level of noise and potential for bias in underlying data, NIST must recommend instances in which even the best de-biasing techniques will be insufficient and data must not be used.

One framework of AI bias suggests that it occurs in layers, with the first layer being bias in the algorithmic models themselves; the second, bias embedded in the data; and the third, bias that emerges from conceptual issues in the development of the models themselves.[78] While NIST's proposal does a good job of laying out biases in the first layer, questions relating to the second and especially the third must be further emphasized.

In conclusion, we ask that NIST meaningfully engage impacted communities in its efforts to reduce bias in AI, analyze both the technical and non-technical factors that contribute to bias, and set clear standards for transparency, data quality, and applications when AI use should be rejected altogether.

Thank you for considering our views.

Sincerely,

American Civil Liberties Union

[76] *See* Stacy Wood, *The Paradox of Police Data*, 2 KULA: Knowledge Creation, Dissemination & Pres. Stud. 9 (2018), https://doi.org/10.5334/kula.34 [https://perma.cc/N2BA-TSFC].
[77] *See* Kristian Lum & William Isaac, *To Predict and Serve?*13 Significance 14 (2016), https://doi.org/10.1111/j.1740-9713.2016.00960.x [https://perma.cc/LY87-995J].
[78] *See* Laurel Eckhouse et al., *Layers of Bias: A Unified Approach for Understanding Problems with Risk Assessment*, 46 Crim. Just. & Behav. 185 (2019), https://doi.org/10.1177/0093854818811379 [https://perma.cc/D47V-ETFB].