Updates
July 20, 2021

NIST Seeks Comment on Proposals to Identify and Manage Bias in Artificial Intelligence

The National Institute of Standards and Technology (NIST) recently published "A Proposal for Identifying and Managing Bias in Artificial Intelligence," a special publication that is part of a series of documents and workshops focused on developing a framework for trustworthy artificial intelligence. NIST, a nonregulatory agency within the U.S. Department of Commerce with a mission to promote U.S. innovation and competitiveness, is accepting comments on its publication until September 10, 2021, and will use them to help shape its planned framework. These initiatives are part of a larger effort by NIST to advance several characteristics needed to cultivate trust in AI systems, including accuracy, explainability, interpretability, privacy, reliability, robustness, safety, and security.

Although AI has demonstrated its potential as a transformative technology, it comes with risks, one of which is harmful bias. As a result, NIST's proposed approach focuses on AI biases that pose harm by causing disparate impacts or excluding certain groups. It presumes that harmful bias is generally present in AI systems and that the challenge lies in identifying, measuring, and managing it.

NIST asserts that harmful biases may be automated within AI systems to perpetuate harms more quickly, extensively, and systematically than would ordinarily occur in day-to-day life. NIST indicated that automated and predictive decisions made within hiring, healthcare, or criminal justice contexts can harm individuals and deepen existing social inequities. While NIST notes that it is unlikely AI technology with "zero risk" can be developed, it is "possible and necessary" to manage and reduce the harmful impacts of bias.

The Challenges Posed by Bias in AI Systems

The challenge of managing bias in AI systems stems from and is exacerbated by their datasets and algorithms. Harmful bias can arise when AI systems are fed data that incompletely or inaccurately captures the real world, or when a developer inappropriately weights certain variables within datasets.

First, NIST notes that AI systems often rely on proxies and indices for their highly complex calculations, but that reliance on these indirect measurements can result in discrimination and performance gaps when they seek to model concepts that are only partially observable by the data. The imprecise correlation between the proxy data and the concepts the AI system seeks to model can result in harmful and discriminatory outcomes. AI systems and their datasets may also over- or under-represent individuals or groups, potentially causing certain populations to be disadvantaged or entirely excluded. Second, even where data properly reflects the real world, such datasets can still exhibit entrenched societal biases or improperly use protected attributes (e.g., gender, age, religion, etc.), further exacerbating the challenge of managing bias in AI systems. Merely excluding protected attributes is insufficient to manage bias, as the attributes can be inadvertently inferred in other ways and thus may still produce negative outcomes for individuals and groups. These problems can be compounded where the algorithmic assumptions of AI systems are not transparent to the public. Third, the decision-making algorithms deployed by AI systems may be untested or unreliable, potentially oversold, or based on questionable or non-existent science, all of which can cause harmful and biased outcomes.

NIST's Three-Stage Approach to Managing Bias

To address this, NIST proposes a three-stage approach that mirrors the three phases of the AI lifecycle: pre-design, design and development, and deployment. Each stage involves distinct activities and stakeholder groups that may introduce bias into AI systems. NIST's proposed approach highlights how different forms of bias—including statistical bias, human cognitive bias, and societal bias—present themselves across each phase and makes practical suggestions for identifying and managing them.

- **Pre-design Stage.** During the pre-design stage, a small number of individuals are responsible for critical product decisions about the purpose of a technology, creating a potential risk that decisions will reflect stakeholders' limited points of view. There is also a risk that stakeholders' optimism or expectations about a product's potential could lead to failures of risk management or to poor communication about potential bias problems. To mitigate these risks, NIST recommends that teams think about the potential societal impacts of products and engage a diverse set of stakeholders early on. Through this wider lens, teams can better recognize problems that may not be apparent to members of the ingroup and predict product applications beyond their initial purposes that could increase the risks associated with bias.
- **Design and Development Stage.** During the design and development stage, software designers, engineers, and data scientists carry out modeling, development, and validation of AI technology. These stakeholders tend to prioritize accuracy, which may not lead to the best outcomes in terms of mitigating harmful bias. Efforts to create an organizational culture in which stakeholders are free to challenge development decisions and are encouraged to document and improve their techniques can lead to improved practices over time. Engagement with subject matter experts, practitioners, and end users during the engagement phase can also help teams refine their approaches and understand how a technology will be used to address potential biases.
- **Deployment Stage.** Finally, during the deployment stage, outside parties begin to interact with the AI technology, causing design decisions that were "poorly or incompletely specified or based on narrow perspectives to be exposed." Additionally, use of the AI technology may vary across social demographics, leading to an unrepresentative sample that can create or further reinforce biases. The real-world applications of AI technology may also drift away from those for which it was initially developed, exacerbating the potential risks of bias. In some cases, this can lead to distrust of the technology; in others, the technology may, inappropriately, be viewed as objective and reinforce disparate outcomes where decisions are offloaded to the tool, or even used to justify users' preexisting biases. Here, deployment monitoring and auditing can help identify and compensate for biases that come into play during this phase.

Existing and Proposed Regulatory Approaches for Addressing Bias in AI Systems

NIST's proposal comes at a time when other regulatory bodies are also focusing on the potential relationship between AI systems and bias. For instance, in April 2021, the Federal Trade Commission (FTC) reiterated current best practices it has publicized for companies to combat algorithmic bias, including to test algorithms "both before [companies] use it and periodically after that—to make sure that it doesn't discriminate on the basis of race, gender, or other protected class." NIST's guidance similarly urges companies to focus on bias across the product development lifecycle, not just at a single point in time. And in 2020, the FTC recommended that companies be transparent in their use of AI, explain their decisions to consumers, ensure their decisions are fair, ensure their data and models are robust and empirically sound, and hold themselves accountable for compliance, ethics, fairness, and non-discrimination. In 2016, the FTC encouraged companies to consider four questions: (1) How representative is the data set? (2) Does your data model account for biases? (3) How accurate are your predictions based on big data? (4) Does your reliance on big data raise ethical or fairness concerns? NIST's

proposed approach raises many similar questions and seeks to offer solutions grounded in the best practices laid out by the FTC.

Other agencies are looking into this issue as well. In April 2021, five financial regulatory agencies sought information and public comments on financial institutions' use of AI, including specific questions aimed at the risk that AI can result in discrimination and how to effectively reduce it. And in late 2020, the U.S. Department of Housing and Urban Development issued a proposed rule that would include certain defenses for companies using algorithms when making housing and credit decisions subject to the Fair Housing Act (FHA). In the final rule, however, the agency decided not to provide a direct defense for algorithms, stating that to do so would be "premature" given expected future developments in the area of algorithms, AI, machine learning, and similar technologies. In addition to these domestic regulatory reviews, as we discussed in a prior update, the European Commission recently issued a proposed regulation that attempts to address the potential risks that AI systems pose to the health, safety, and fundamental rights of Europeans caused by AI systems.

Clearly, mitigating bias in AI systems remains top-of-mind for regulators across the United States, and NIST's proposed approach could provide a framework that helps companies to address this challenge going forward.

Key Takeaways for Companies Developing AI Systems

While AI-specific regulations are still few and far between, a number of state and federal laws prohibit discrimination on the basis of a protected class in specific scenarios. A 2014 White House Study found that it is rare for algorithmic bias to arise from intentional disparate treatment; accordingly, these laws tend to apply regardless of companies' intent. For example, Section 5 of the FTC Act generally prohibits unfair or deceptive practices, which the agency notes would include the sale or use of racially biased algorithms. Companies may also be exposed to liability under sector-specific laws. In the financial sector, the Equal Credit Opportunity Act (ECOA) makes it illegal for a company to use a biased algorithm that results in credit discrimination on the basis of a protected class. The Fair Credit Reporting Act (FCRA) also imposes disclosure requirements that could be violated if covered entities base credit decisions on algorithms that are implicitly biased and fail to properly disclose the bases for such decisions to consumers. Within the healthcare industry, Section 1557 of the Affordable Care Act (ACA) prevents any healthcare provider receiving federal funds from refusing to treat or otherwise discriminating against an individual based on protected classes. As NIST's proposal shows, harmful biases may embed themselves early in the product development life cycle. If parties wait until a product is deployed to consider and address the effects of harmful bias, it may be more costly or even too late to address its root causes effectively.

It is also in companies' business interests to develop a holistic approach to mitigating harmful bias. Unless companies developing AI technology for the public can demonstrate that they are actively engaged in identifying and mitigating bias, the mere perception that AI systems may be biased can have a chilling effect on adoption of the technology. Public suspicions about the prevalence of bias within AI systems, even among those who are ultimately unharmed by it, could curb the technology's ability to revolutionize higher-risk areas such as finance and health.

Conversely, if such companies are seen as using technology to add value to people's lives, while reducing disparate impacts and other harmful effects of bias, a virtuous circle may begin, as the public becomes more supportive and trusting of the technology. As Dr. Sian Townson observed in the Harvard Business Review, "[r]educing bias is not just a socially responsible pursuit—it also makes for more profitable business. The early movers in reducing bias through AI will have a real competitive advantage on top of doing their moral duty."

Doing so effectively will likely require close collaboration between business, legal, and product teams across the product development life cycle. NIST offers one approach for doing this, which can provide a helpful starting place for businesses.

For more information about AI in healthcare, see Marc Martin et al., Combating Bias in Artificial Intelligence and Machine Learning Used in Healthcare (Sept. 2020).

The authors recognize the contributions of summer associates Stephanie Duchesneau and Sam Klein to this update.

Authors

Marc S. Martin

Partner
MMartin@perkinscoie.com 202.654.6351

Explore more in

Technology Transactions & Privacy Law Communications Advertising, Marketing & Promotions Artificial Intelligence & Machine Learning

Related insights

Update

'Tis the Season… for Cybercriminals: A Holiday Reminder for Retailers

Update

Employers and Immigration Under Trump: What You Need To Know