INSIGHT 4

BIAS IN AI SYSTEMS COULD REMAIN FOREVER

Bias is a feature of both human and AI decision making. As the data used to train AI is often biased in hard-to-fix ways, growing reliance on AI in decisionmaking systems could spread bias and lead to significant harm. Bias may never be eliminated, in part due to conflicting perspectives on fairness.

TODAY

Bias in AI is seen as a major issue capable of automating discrimination at scale in ways that can be difficult to identify. While human decisions are also biased, one of the major risks of automating high-stakes decisions is that these become more widespread and less detectable, increasing the possibility of systemic errors and harms. While a single biased manager could decide to give higher interview scores to the few job applicants that look and speak like them, a biased AI model could have a similar effect on potentially thousands of people across organizations, sectors, or countries.

Many AI products claim to be less biased than human decision makers but independent investigations have revealed systematic failures and rejections.[1] For example, an audit of two AI hiring tools found that the personality types it predicted varied depending on whether an applicant submitted their CV in Word or raw text.[2] Similar tools have discriminated against women[3] or people with disabilities.[4] Bias is embedded in AI in many parts of its lifecycle — training data, algorithmic development, user interaction, and feedback.[5]

Bias may be impossible to eliminate because the data used for training AI models is itself often biased in ways that cannot easily be fixed. Controlling results can also cause problems. For example, an AI model that learns to discard racially sensitive wording might omit important information about the Holocaust or slavery.[6] Further, algorithms often cannot compute different notions of fairness at the same time, leading to constantly different results for certain groups.[7, 8, 9]

FUTURES

In a future where bias can never be eliminated — whether human or algorithmic — societies may need to rethink current ideas about fairness and how to best achieve it. People do not necessarily agree on the meaning of "fair." For example, some consider affirmative action to be fair while others do not. Institutions could adopt standards intended to distribute resources — jobs, grants, awards, or other goods — in ways that explicitly attempt to repair historical injustices. Organizations seeking to avoid systemic bias may use an "algorithmic pluralism" approach, which involves various elements in the decision-making process and ensures no algorithms severely limit opportunity.[10]

Efforts could be made to reduce bias in AI systems to an acceptable level, though eliminating it entirely could be impossible. Pushback may continue against using AI technologies in certain sensitive domains, such as policing or hiring. Alternatively, these technologies could continue to improve and become less biased in the future. Either way, there will likely be a continued push to reducing bias in AI technologies.

IMPLICATIONS

- **Systemic harms or failures could become institutionalized** in contexts where single algorithms are allowed to make bulk decisions about people's access to certain resources (e.g. jobs, loans, visas)
- **Human biases could become greater** among those who use AI systems, as people learn from and replicate skewed AI perspectives, carrying bias with them beyond their interactions
- **Disagreements about the best ways to code for algorithmic fairness** may result from different definitions of what fairness actually means. This could lead to completely different results for similar technologies or systems
- The inability to eliminate bias from algorithms could ultimately **lead to political, social, or economic divisions**
- If decisions become more distributed, including various algorithms and humans at different points in a process, **it could be difficult to make discrimination claims** or identify a responsible party for discrimination
- High-profile cases of algorithmic discrimination could lead to **loss of trust in AI decision-making systems**, particularly in policing and healthcare, and an increase in litigation

Endnotes

1 Helhoski , Anna. AI Could Prevent Hiring Bias — Unless It Makes It Worse. NerdWallet, 12 June 2023.

2 Rhea, Alene K., Kelsey Markey, Lauren D'Arinzo, Hilke Schellmann, Mona Sloane, Paul Squires, Falaah Arif Khan, and Julia Stoyanovich. An external stability audit framework to test the validity of personality prediction in AI hiring. Data Mining and Knowledge Discovery 36, no. 6 (2022): 2153-2193.

3 Andrew, Lori, and Hannah Bucher. Automating Discrimination: AI Hiring Practices and Gender Inequality. Cardozo Law Review. Accessed 15 August 2024.

4 Brown, Lydia, Ridhi Shetty and Michelle Richardson. Algorithm-Driven Hiring Tools: Innovative Recruitment or Expedited Disability Discrimination?, 3 December 2020.

5 Samuel, Sigal. Why It's so Damn Hard to Make AI Fair and Unbiased. Vox, 19 April 2022.

6 Ferrara, Emilio. Eliminating Bias in AI May Be Impossible – a Computer Scientist Explains How to Tame It Instead. The Conversation, 19 July 2023.

7 Kleinberg, Jon. Inherent Trade-Offs in Algorithmic Fairness. YouTube, 10 April 2018.

8 Dwork, Cynthia. The Emerging Theory of Algorithmic Fairness. YouTube, 6 September 2018.

9 Raghavan, Manish. What Should We Do When Our Ideas of Fairness Conflict? Communications of the ACM 67, no. 1 (January 2024): 88–97.

10 Jain, Shomik, Vinith Suriyakumar, Kathleen Creel, and Ashia Wilson. Algorithmic Pluralism: A Structural Approach To Equal Opportunity. arXiv, 21 September 2023.

PDF: PH4-214/2025E-PDF
ISBN: 978-0-660-76887-8

Aussi disponible en français sous le titre : Les biais des systèmes d'IA pourraient perdurer.