Consilio Institute: Article

BIAS IN AI: UNDERSTANDING AND MITIGATING THE RISK

by Xavier Diokno
Senior Director, Innovation Solutions
by Maureen O'Neill
Senior Vice President, Strategic Client Experience

ADVANCED LEARNING INSTITUTE

BIAS IN AI: UNDERSTANDING AND MITIGATING THE RISK

The world is increasingly turning to artificial intelligence (AI) to help us make decisions that impact business processes, financial systems, and our personal health and welfare. While we strive to build AI tools that generate objective, data-driven outputs, the potential for these systems to create biased outcomes has garnered significant attention. There is an inherent risk that the algorithms used by AI systems can perpetuate, and even amplify, societal biases and inequities based on race, ethnicity, gender, socio-economic status, sexual orientation, and other characteristics.

Biased AI systems can have profound consequences. In the employment context, gender bias embedded in an AI hiring tool can prevent women from being hired. In financial services, biased credit-scoring algorithms can deny loans or set unfavorable terms for historically marginalized groups. In healthcare, a biased AI model might misdiagnose or under-treat certain demographic groups, resulting in unequal health outcomes.

As AI and DEI leaders at Consilio, we thought it would be useful to provide a short overview of how bias unintentionally creeps into AI systems, and how the designers and users of these systems can mitigate the risk of bias.

Sources of Bias in AI

- **Historically Inequitable Data and Temporal Data "Drift."** A significant contributor to bias in AI is the very data on which the systems are trained. If the training data reflects historical inequities, these patterns are likely to be embedded in the AI's outcomes. For instance, an AI model trained on hiring data from a company that historically hired fewer women might learn to favor male candidates, thereby perpetuating gender discrimination. Also, the data used to train an AI model can become out of date, "drifting" away from an understanding of the present-day world. Some AI models effectively have a view of the world that's frozen in time, which can bias the results through a lack of knowledge of more current events, outdated views on certain topics, and lack of understanding about emerging viewpoints.
- **Human Biases.** Biased decisions made by the people developing AI systems may also infiltrate AI during the design phase. Developers may, consciously or unconsciously, make choices that embed their own biases. For example, they might select decision-making criteria that are easy to measure, such as zip code, which may inadvertently introduce racial or socioeconomic biases.
- **Toxicity.** AI models that are trained on the content of the internet will contain a lot of morally dicey information—aka "toxicity." These systems have ingested racist and misogynistic views, information about how to create nuclear bombs, false stories about immigrants eating pets, instructions for creating malware, and so on. Although developers of AI models strive to ensure that the systems give harmless and helpful responses, this information is still buried within the parameters and can be expressed through inherently biased outputs.
- **"Noteworthy" Bias.** AI models trained on the internet will represent the distribution of information found there. Generally, people write about events and topics that are noteworthy, rather than those that are mundane, and this tendency can bias AI output. For example, a prompt to an AI system to "write a news headline about Springfield, Ohio" will more likely suggest "Backlash grows over baseless claims that immigrants are eating local pets," than "First snow of the season falls." This skews AI output towards more sensational topics, which could introduce biases on societal issues that fall on the extremes of the spectrums.

- **Feedback Loops.** Finally, AI algorithms themselves can generate bias through feedback loops. In predictive policing, for example, an AI system might focus on areas with high crime reports based on previous policing patterns, leading to increased police presence and further arrests in those areas, reinforcing the AI's biased assumptions.

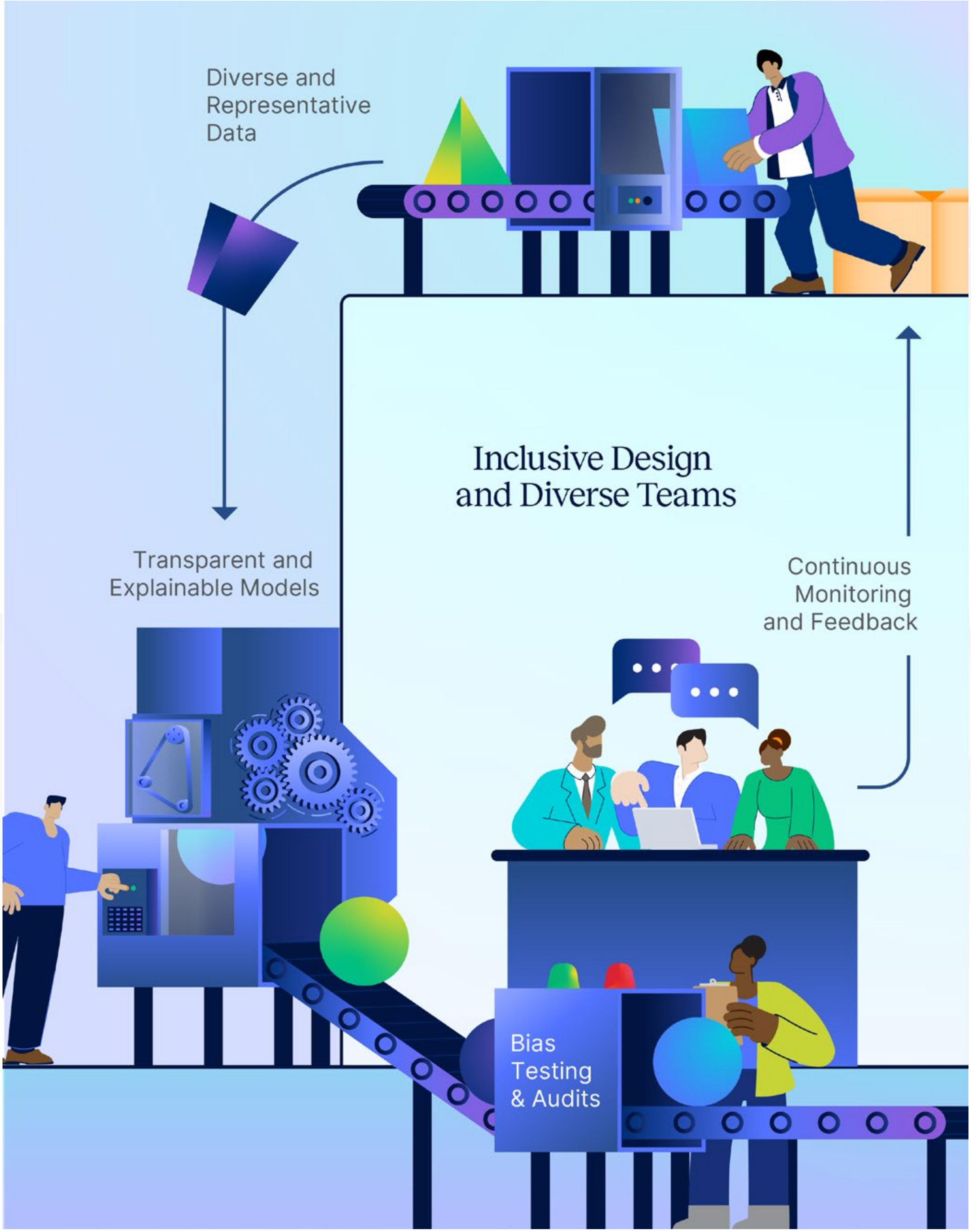
Mitigating Bias in AI

The risk of bias in AI systems is not inevitable, and there are a number of measures that can reduce the risk.

- **Diverse and Representative Data.** One of the most effective ways to reduce bias is to ensure that training data are drawn from a diverse array of sources, and the data are representative of the population the AI will serve. Regular audits of datasets can help identify and rectify any biases within them. Techniques such as oversampling underrepresented groups can also create a more balanced dataset.
- **Transparent and Explainable Models.** Using interpretable models and/or incorporating transparency mechanisms in AI systems helps stakeholders understand how decisions are made, making it easier to spot and correct biases. Techniques like model explainability can reveal patterns that may indicate bias, allowing developers to make necessary adjustments.
- **Bias Testing and Audits.** Regular testing for bias throughout an AI system's lifecycle is essential. Bias testing tools can reveal discrepancies in outcomes for different demographic groups, allowing developers to adjust models as needed. Independent audits by third-party organizations can also help provide an objective view of the AI's fairness and performance.
- **Inclusive Design and Diverse Teams.** Involving diverse voices in the AI development process can help identify potential biases that could be overlooked by a homogenous team. Recognizing bias is often a matter of perspective, and people from different backgrounds might notice different biases. Including individuals with diverse experiences, expertise, and training allows for a more comprehensive approach to anticipating and mitigating biases.
- **Continuous Monitoring and Feedback.** AI systems should be continuously monitored in real-world applications to detect any emerging biases. Feedback mechanisms can also allow users to report issues, which can be used to improve the system over time.

Bias in AI is not merely a technical issue but a societal one, as biased AI can perpetuate existing inequalities and create new forms of discrimination. By focusing on data quality, transparency, continuous bias audits, and inclusive development practices, organizations can significantly reduce bias in their AI systems. As AI becomes more pervasive, developing fair and unbiased algorithms will be essential to ensuring that technology serves everyone equitably and ethically.

Processes to Mitigate Bias in AI


Diverse and Representative Data
Inclusive Design and Diverse Teams
Transparent and Explainable Models
Continuous Monitoring and Feedback
Bias Testing & Audits

ABOUT THE AUTHOR

Xavier Diokno is a Senior Director at Consilio, a global leader in eDiscovery, document review and legal consulting services. Xavier has a bachelor's degree in computer science from Southern Illinois University, a master's degree in computer science from the University of Illinois at Chicago and a juris doctor degree from DePaul University College of Law. He is licensed to practice in the state of Illinois and the United States Patent and Trademark Office.

Prior to becoming an attorney, Xavier worked in the information technology industry for ten years in database administration and software development. For more than a decade, Xavier was part of Consilio's Data Analytics group, where he oversaw the team's tripling in size, as well as numerous large-scale projects involving Technology-Assisted Review, Immediate Case Assessments™, and novel analytics research. Xavier now applies his technical and legal experience to overseeing Consilio's Innovation initiatives, including researching new technologies like artificial intelligence and developing their application to legal services.

Xavier Diokno

Senior Director, Innovation Solutions

e xdiokno@consilio.com

consilio.com

ABOUT THE AUTHOR

Maureen O'Neill serves as Consilio's Diversity, Equity & Inclusion Officer. She leads the development and implementation of the company's strategies, policies, and programs for ensuring a diverse, inclusive, and equitable workplace. Maureen guides the organization in carrying out its mission of attracting and encouraging a diversity of backgrounds, experiences, and ideas, and fostering an inclusive culture among its global workforce.

Prior to joining Consilio, Maureen was a partner at Paul Hastings LLP, where she represented Fortune 100 and other multinational companies in complex employment litigation, including class and collective actions. Maureen also co-chaired the Firm's E-Discovery Practice Group. In that role, she advised clients on various e-discovery issues, and consulted internally with attorneys at the Firm on best practices for litigators engaged in e-discovery.

Maureen O'Neill

Senior Vice President,
Strategic Client Experience

e maureen.oneill@consilio.com

consilio.com