# IBM Policy Lab: Mitigating Bias in Artificial Intelligence

*Anjelica Dortch, Technology Policy Executive, IBM Government & Regulatory Affairs*
*Dr. Stacy Hobson, Director, Responsible and Inclusive Technologies, IBM Research*

There's no question that human biases could influence algorithms and result in discriminatory outcomes. However, it is difficult to discern how pervasive these biases are in the technology we develop and use in our everyday life. While mitigation of bias in AI models might be challenging for some AI and automated decision-making systems, it is imperative to reduce the likelihood of negative outcomes.

Our society continues to evolve with rapid innovation in emerging technologies, in particular AI. Industry, academia, governments, and consumers have a shared responsibility to ensure that Al systems are tested and assessed for bias. Furthermore, any action or practice prohibited by anti-discrimination laws should continue to be prohibited when it involves an automated decision-making system. To support bias mitigation strategies, organizations should work to create, implement, and operationalize AI ethics principles, and ensure appropriate governance is in place to provide ongoing review and oversight of AI systems.

Without the right safeguards, AI could cause harm and exacerbate existing inequalities. At IBM, we believe that harnessing the transformative potential of AI requires a commitment to actively develop and use it responsibly to prevent discriminatory outcomes that could negatively harm individuals and their families. One critical aspect of the responsible development of AI is the focus on identifying and mitigating bias. In recent years, IBM has shared research findings and made available tools to support bias mitigation and provide companies and their consumers with a better understanding of the AI systems they build and use every day. These include the AI Fairness 360[1] toolkit, AI FactSheets[2], IBM Watson OpenScale[3], and new IBM Watson capabilities designed to help businesses build trustworthy AI.

Last year, the IBM Policy Lab called for "precision regulation" to strengthen trust in AI with a risk-based AI governance policy framework based on accountability, transparency, fairness and security

that called on industry and governments to take actions. In light of how the public dialogue around AI bias has evolved, that perspective – applying narrowly-tailored policy approaches to addressing discrete harms – is more important than ever. That is why, in response to renewed attention to inequalities and the way that technology – in areas like criminal justice, financial services, health care, and human resources – can be misused to exacerbate injustice towards marginalized groups, IBM is advocating that policymakers take additional steps to shape a legislative environment that is conducive to addressing legitimate societal concerns.

IBM is committed to advocating for diversity, equity, and inclusion in our society, economy, and the technology we build. As such, we are calling on governments to implement five policy priorities to strengthen adoption of testing, assessment, and mitigation strategies to minimize instances of bias in AI systems:

1. **Strengthen AI literacy throughout society.** Greater understanding across society on what AI is, its potential benefits, and how to engage with AI systems could accelerate its growth and ability to be trusted. Promoting AI literacy would arm more of society with the skills needed to adapt to a changing world where AI will be prevalent.

    *a.* The development and implementation of a national "AI for All" agenda could foster a more inclusive and diverse AI ecosystem and support eradicating general misconceptions around AI.

    *b.* Increasing investment in education at all levels to build AI into the curriculum and directing greater research funding for AI testbeds with minority-serving institutions at the table, could help to ensure that in the future a more diverse range of stakeholders guide the design, development and application of AI systems. While the AI field today may not reflect the demographics of our society, moving toward a more diverse AI ecosystem—from developers to users—could enable organizations to avoid and mitigate unwanted AI bias by including and reflecting the interests and values of communities likely to be impacted.

    *c.* Science and technology agencies or ministries should also prioritize partnership opportunities that advance racial equity in AI. Where possible, these collaborations should include representatives from communities most impacted by inequities.

2. **Require assessments and testing for high-risk AI systems.** While all entities developing and owning high-risk AI systems should assess and test their systems, any mandatory requirements should focus on protecting consumers from the greatest harm, while enabling innovation. This means:

    *a.* Requiring bias testing and bias mitigation, in a robust and transparent manner, for certain high-risk AI systems such as law enforcement use cases. These high-risk AI systems should also be continually monitored and re-tested;

    *b.* Focusing any requirements for conducting impact assessment prior to deployment on owners of those high-risk AI systems that pose the greatest potential to harm;

    *c.* Documenting the assessment processes in detail, making them auditable, and retaining them for a minimum period of time;

    *d.* Convening and driving national and international forums to accelerate consensus around clear and consistent standards, definitions, benchmarks, frameworks, and best practices for trustworthy AI;

*e.* Providing resources and expertise to help all organizations—not just large corporations—ensure their AI is deployed responsibly;

*f.* Increasing investment in research and development around bias testing and mitigation to ensure leading scientific approaches are targeted at mitigating bias; and

*g.* Supporting accelerated developer training around bias to ensure appropriate training aimed at understanding and mitigating biases and recognizing how bias could be unintentionally introduced into AI systems during the development pipeline.

3. **Require AI transparency through disclosure.** Developers and owners should disclose to users when they are interacting with AI technologies with little or no human involvement. Furthermore, disclosure to users should also occur when a high-risk AI system is used to make decisions, similar to the disclosure requirements outlined in the Fair Credit Reporting Act (FCRA)[4] and General Data Protection Regulation (GDPR)[5]. And in the case of these automated decision-making systems, at a minimum, the disclosure should communicate to the user why and how a particular decision was made using AI.

4. **Require mechanisms for consumer insight and feedback.** Similar to many consumer safety guidelines, operators of high-risk applications should include with their disclosure a communication mechanism (e.g., email, phone number, or mailing address) to capture issues, concerns, or complaints from consumers related to automated decisions. Owners should act responsibly by conducting ongoing reviews of consumer concerns, and when appropriate, work to address systemic issues.

5. **Establish universal use limitations of AI and adopt responsible licensing practices.** To prevent high-risk AI systems from being leveraged for prohibited, irresponsible, and harmful uses, we call for:

*a.* the establishment of universal use limitations in high-risk AI applications to prohibit the use of AI for mass surveillance, racial profiling, and violations of basic human rights and freedoms; and

*b.* expanding the development, education, and adoption of responsible AI licensing terms for open-sourced AI software and applications. This voluntary licensing framework can be used by individual developers and organizations to include clauses for limiting the use of potentially harmful AI systems.

New laws, regulatory frameworks, and guidance for mitigating bias in AI systems are on the horizon. If well-crafted with these priorities as a foundation, these measures can provide industry and organizations with clear testing, assessment, mitigation and education requirements to enhance consumer confidence and trust in AI.

IBM stands ready to work together with lawmakers to act on these imperatives and ensure that the benefits of this incredibly promising technology are felt broadly across society.

---

[1] http://aif360.mybluemix.net/

[2] https://aifs360.mybluemix.net/introduction

[3] https://www.ibm.com/cloud/watson-openscale

[4] The Fair Credit Reporting Act (FCRA), 15 U.S.C. § 1681

[5] EU General Data Protection Regulation (GDPR): Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016