Category

Data & AI

AI Bias Detection, Mitigation & Fairness Playbook

Testing for bias across the groups that actually matter, not just overall accuracy

- Practitioner
- Advanced
- Template Included

- Administrator
- 29 January 2020
- 3 minute(s)

Overview

A framework for AI bias detection, mitigation, and fairness — disaggregated testing across relevant demographic and use-case groups, understanding different fairness definitions and their trade-offs, and building ongoing bias monitoring rather than treating fairness as a one-time pre-deployment check.

If an AI model shows strong overall accuracy, is that sufficient

evidence it's not biased? No — overall accuracy can mask significantly worse performance for specific demographic or use-case groups, which is exactly why disaggregated testing across relevant groups, not just aggregate accuracy, is essential for genuine bias detection.

Why do different fairness definitions matter, rather than there

being one clear standard for "fair" AI? Different mathematical fairness definitions (equal accuracy across groups, equal false-positive rates, equal outcome rates) can actually conflict with each other in certain situations, meaning organizations need to make deliberate, informed choices about which fairness definition matters most for their specific application, rather than assuming a single universal standard exists.

{ "@context": "https://schema.org", "@type": "FAQPage", "mainEntity": [{ "@type": "Question", "name": "If an AI model shows strong overall accuracy, is that sufficient", "acceptedAnswer": { "@type": "Answer", "text": "evidence it's not biased?\nNo — overall accuracy can mask significantly worse performance for\nspecific demographic or use-case groups, which is exactly why\ndisaggregated testing across relevant groups, not just aggregate\naccuracy, is essential for genuine bias detection." } }, { "@type": "Question", "name": "Why do different fairness definitions matter, rather than there",

"acceptedAnswer": { "@type": "Answer", "text": "being one clear standard for \"fair\" AI?\nDifferent mathematical fairness definitions (equal accuracy across\ngroups, equal false-positive rates, equal outcome rates) can actually\nconflict with each other in certain situations, meaning organizations\nneed to make deliberate, informed choices about which fairness\ndefinition matters most for their specific application, rather than\nassuming a single universal standard exists." } }] }

Subscriber access

Unlock this playbook

This playbook — including every framework, template, and step-by-step section — is available free to Think Insights subscribers. Enter your email to unlock it instantly and get our weekly insights newsletter. No account needed, and access is remembered on this device.

First Name

Last Name

Email

Country - Select -

■ I agree to data processing in accordance with GDPR

Submit

Leave this field blank

References

Related playbooks

Data & AI

Data Literacy for Multi-Disciplinary Data & AI Communities of Practice Playbook

A framework for building and sustaining multi-disciplinary data and AI communities of practice — bridging genuinely different literacy levels and functional per

- Practitioner
- Intermediate
- Template Included

Read playbook

Data & AI

Measuring Data Literacy Maturity & Impact Playbook

A framework for measuring data literacy program maturity and genuine business impact — behavioral change indicators, decision-quality outcomes, and maturity sta

- Practitioner
- Intermediate
- Template Included

Read playbook

Data & AI

Data Literacy for Risk Assessment & Controls Design Playbook

A framework for building data literacy specifically for risk assessment and internal controls design — data-driven risk prioritization, control effectiveness me

- Practitioner
- Intermediate
- Template Included

Read playbook

Talk to us about AI bias and fairness

Tags

- AI Bias
- AI Fairness
- Bias Mitigation

Author

Administrator

Think Insights Administrator