Catch Me If You Can: The Dynamic Nature of Bias in ML

Tarafdar et al. (2026)

Context

Most organisations treat algorithmic bias as a fixed flaw to be detected and removed once. This study of a US AI hiring firm shows why that assumption fails. Through 27 interviews across six departments, the authors trace how bias moves through an ML system's lifecycle. As each mitigation action reshapes the conditions around it, fixing bias in one place quietly seeds it elsewhere, meaning bias behaves more like a moving target that one-off audits cannot pin down.

Key Insights

Bias in hiring is dynamic, not static

- Bias emerges from three interacting sources: training data, the model, and how humans use outputs
- A source of bias in one stage cascades into later stages, so when it surfaces its origin is hard to trace
- Mitigation actions can reduce the targeted bias, but often introduce a new bias source elsewhere

Mitigation is a cross-functional, ongoing job

- Technical checks alone (e.g. the US 4/5th hiring rule) miss ripple effects, so measurement & bias mitigation must be cross-functional (data science, psychology, client-facing roles) to intervene across stages
- A dedicated bias governance owner with a holistic view should decide & monitor where audits are needed

Bias mitigation action	**Bias-decreasing effect** (intended "control")	**Bias-increasing effect** (side effect "drift")	**FORGOOD tension** (not in original source)
Validate competencies against the O*NET database and questions against the literature	Assessments designed with correct competencies, reducing measurement bias	Minority demographics not adequately represented in the literature and all jobs not in O*NET, so competencies for minority groups are unaccounted	**Fairness**: the benchmark itself disadvantages under-represented groups
Collaborate with clients to assess the ML application's performance	Clients identify & rectify bias in their hiring	Client walks away from future bias analysis, increasing bias in their recruitment process	**Delegation**: the organisation lacks leverage to compel ongoing scrutiny
Establish an adverse impact ratio for parameters (e.g. race)	Model outputs confined within the ratio, reducing bias from the applicant-pool vs training-pool gap	The ratio does not account for increasingly varied interracial skin tones, a temporal change in the parameter	**Openness**: a fixed metric hides drift no longer visible in the numbers

Implications

- Where in your own ML decision systems could a bias fix quietly seed a new source of bias?
- What would change in your monitoring if fairness were treated as a continuous maintenance cost?
- Who would need to hold a cross-functional, cross-stage governing view of bias?

Tarafdar, M., Rets, I., & Hu, Y. (2026). Catch Me If You Can: The Dynamic Nature of Bias in Machine Learning Applications. Information Systems Journal 1–33. https://doi.org/10.1111/isj.70034