AI Can See Bias. It Still Acts On It.

New benchmark reveals a hidden flaw in how LLMs decide vs. describe.

Swipe for insights →

THE DATA

+0.55

Normative-Descriptive Separation (NDS)

Claude Sonnet 4.6 shows the strongest ability to distinguish between describing and prescribing outcomes.

This score measures how well a model separates normative reasoning (what should happen) from descriptive reasoning (what is likely to happen). A positive NDS indicates that the model adjusts its output based on whether it's being asked to predict or to decide. The higher the score, the more reliably the model maintains ethical separation in its responses.



Source: arXiv · Published November 23, 2026
https://arxiv.org/abs/2611.26870

01

Decision Framing Triggers Hidden Bias

The study reveals that even when models are trained to avoid demographic discrimination, they still exhibit strong bias when asked to 'decide' rather than 'describe' outcomes. This suggests that the act of decision-making amplifies latent biases, even in models designed for fairness. The difference in behavior under different framing is not just a side effect — it's a core design vulnerability in current AI systems.



02

Bias Is Not Just in Data — It’s in Design

The fact that reversing candidate order leads models to favor the first option — even when candidates are identical — exposes a fundamental flaw in how AI systems process choice. This isn't just about training data; it's about architectural defaults. The model’s behavior under decision framing reflects a systemic tendency to prioritize order, which undermines the integrity of any fairness benchmark that doesn't account for framing effects.

George Arnold Alexander III, MPA

Partner, ALEX Companies
Associate Broker, Alexander Tiffany Southwest, LLC
Licensed Broker in TX, CO, NC, AL

george@alex.realestate · +1 713.918.9951

ALEX

ALEX Intelligence
deep intelligence

ALEX Real Estate
genius real estate

ALEX Advisory
uncontested strategy

ALEX Cares
purposeful philanthropy