What we think is missing in the fairness literature

- Clean, random (experimental) variation in **programming practices**.
- Paired with clear **outcome measures** of success/failure.
- So that the research community can *causally* link programming practices with the presence (or absence) of bias in code.
- ... and link these results back to theory.

This Paper: Field Experiment in AI Development

- $\approx$ 400 programmers
- Same task:
  - Predict performance on a standardized math test
  - For 20K randomly selected people (using administrative data).
  - Using over 5000 covariates/person.
- Under four randomized experimental conditions.

Preview of Results (I): Interventions

- Positive Result: Non-technical reminders
    - Very effective.
    - About 60% of benchmark #1 (completely unbiased data).
- **Null Result**: Incentives
    - Affected effort (programming hours)
    - ... but *not outcomes*.
- Negative Result: Technical advice reversed the benefit of the reminder.
    - i.e., it made algorithmic bias worse.

Preview of Results (II): Programmer Characteristics

- Broadly uncorrelated with bias in code.
  - True for demographics.
  - As well as for implicit association test (IAT).
- However, **prediction errors** are correlated within demographics.
- This implies bias reduction through cross-demographic averaging.