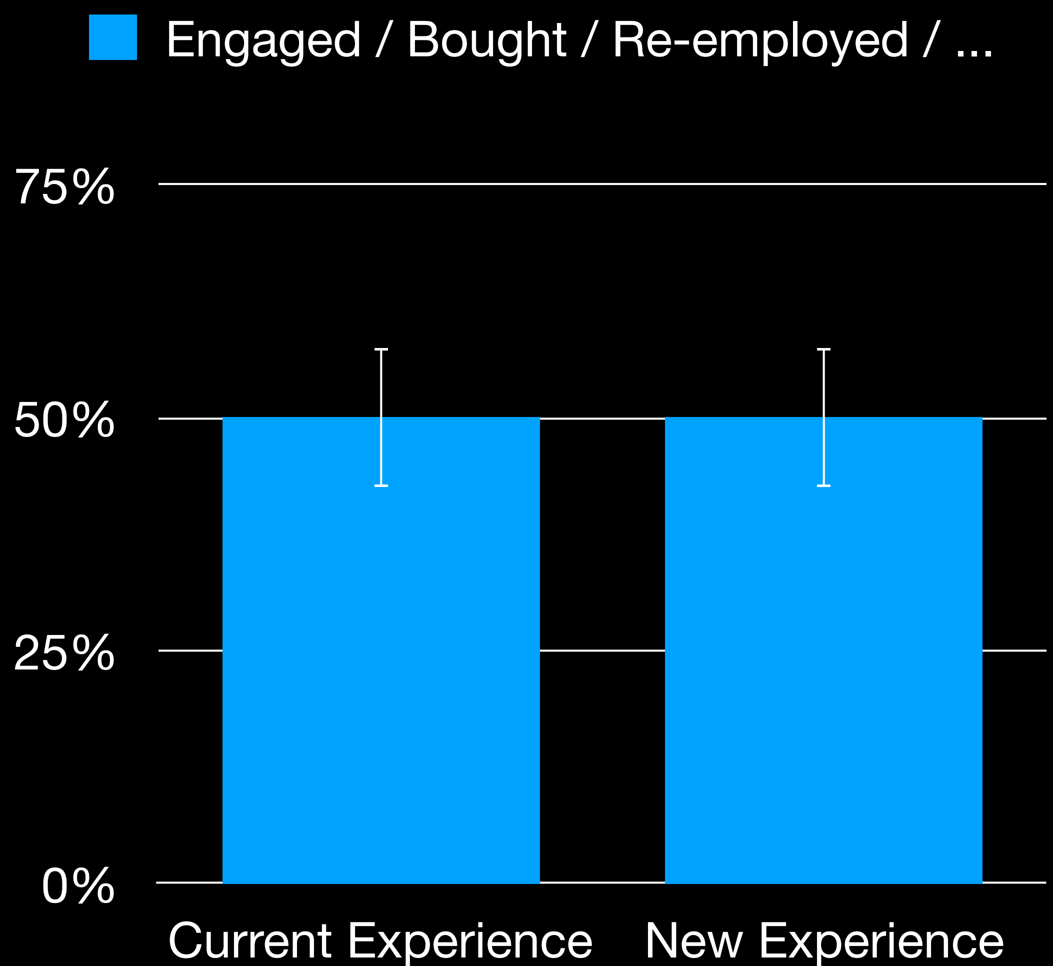
What's the Harm?

Sharp Bounds on the Fraction Negatively Affected by Treatment

Nathan Kallus

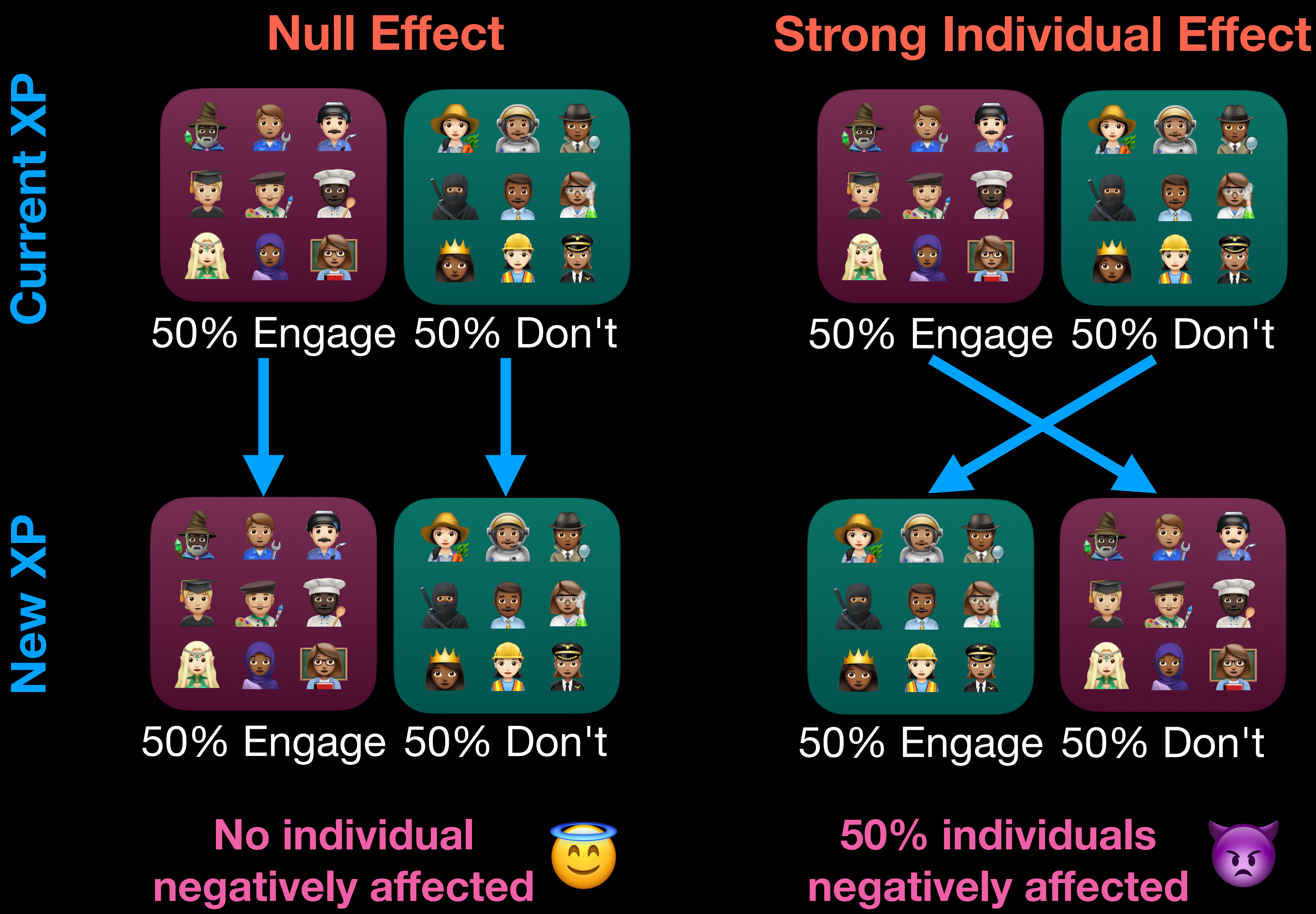
Cornell & Netflix

A/B test a proposed change



Looks totally harmless 😇 Is it tho? 👿

Two equally possible scenarios



So... What's the Harm?

- Fraction Negatively Affected: $\mathrm{FNA} = \mathbb{P}(Y(1) < Y(0))$
  - Crucial for judging a change's impact on downstream behavior, fairness, operations
- Unlike $\mathrm{ATE} = \mathbb{E}[Y(1) - Y(0)]$, FNA is *not* identifiable
  - No amount of data, even if experimental, will allow us to pin FNA down
- Can still hope to *partially* identify, *i.e.*, give bounds
  - But want *informative* bounds
  - *i.e.*, not [0%, 50%]

This paper

- **Sharp bounds** (i.e., tightest possible) on FNA with covariate information on units
  - Also bounds on related quantities

- **Estimation & inference** on bounds, which involve complex functions like the conditional avg treatment effect (CATE)
  - **Locally robust**: fast convergence rates and calibrated confidence intervals even when these functions are estimated slowly by ML blackboxes
  - **Doubly valid**: even if CATE (& similar) is misspecified, get 2 chances at valid (albeit conservative) bounds

- So, gives credible inference, can support addressing harm:
  - Focusing on bounds accounts for ***unknowables***
  - Robustness ensure reliability under ***estimation errors***