Text as Causal Mediators

Research Design for Causal Estimates of Differential Treatment of Social Groups via Language Aspects

We propose a research design for (observational) causal estimates of the effects of social group signals on speakers' responses via language as causal mediators.

Identification assumptions

- Sequential ignorability (Imai et al, 2010)

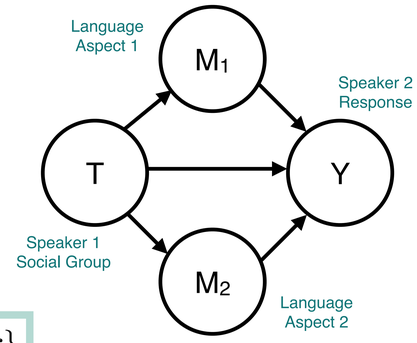
$$\{Y_i(t', m), M_i(t)\} \perp\!\!\!\perp T_i \mid X_i = x$$

$$Y_i(t', m) \perp\!\!\!\perp M_i(t) \mid \{T_i = t, X_i = x\}$$

- Mediator Independence

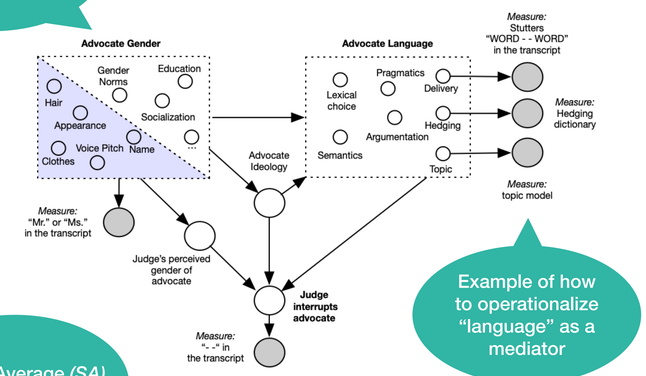
$$\forall j, j' : \; M_i^j(t) \perp\!\!\!\perp M_i^{j'}(t) \mid \{T_i = t, X_i = x\}$$

Simplified causal graph



Expanded causal graph

Addresses "social group" as a causal treatment



Example of how to operationalize "language" as a mediator

Estimation

Sample Average *(SA)* for each mediator *j*

$$\text{SA-NDE}^j = \frac{1}{N}\sum_{i=1}^{N}\sum_{x\in\mathcal{X}}\sum_{m\in\mathcal{M}^j}\Big(\hat{f}^j(Y|M_i^j = m, T_i = 1, X_i = x) - \hat{f}^j(Y|M_i^j = m, T_i = 0, X_i = x)\Big)\hat{g}^j(m|T_i = 0, X_i = x)$$

$$\text{SA-NIE}^j = \frac{1}{N}\sum_{i=1}^{N}\sum_{x\in\mathcal{X}}\sum_{m\in\mathcal{M}^j}\hat{f}^j(Y|M_i^j = m, T_i = 0, X_i = x)\Big(\hat{g}^j(m|T_i = 1, X_i = x) - \hat{g}^j(m|T_i = 0, X_i = x)\Big)$$

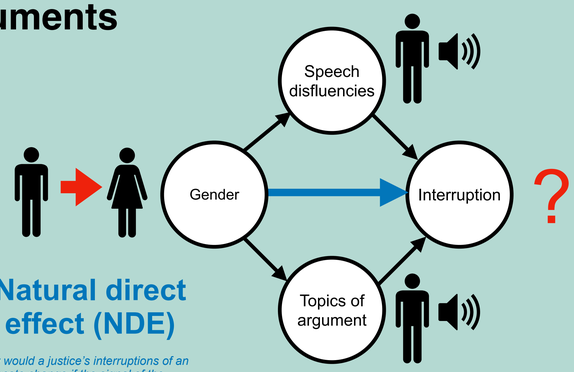
- Fit *f*, *g* on train set. Then at inference time, apply the fitted models to real confounders from the test/inference set and "counterfactual" treatment and mediator values.
- We will also need models of mediators given raw text. An open question is whether to infer this separately or jointly with *f* and *g*.

Theoretical case study from U.S. Supreme Court oral arguments

(Photo Credit: Fred Schilling, Collection of the Supreme Court of the United States)

Q: Why do some justices interrupt female advocates more than male advocates?

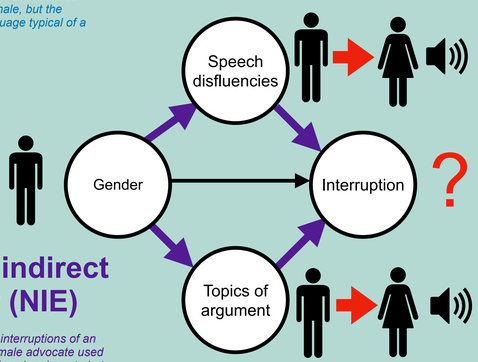
Hypothesis 1: Implicit gender bias



Natural direct effect (NDE)

How would a justice's interruptions of an advocate change if the signal of the advocate's gender the justice received flipped from male to female, but the advocate still used language typical of a male advocate?

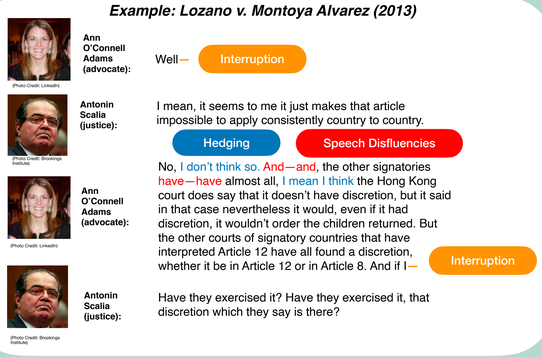
Hypothesis 2: Women are "less effective" speakers



Natural indirect effect (NIE)

How would a justice's interruptions of an advocate change if a male advocate used language typical of a female advocate but the signal of the advocate's gender the justice received remained male?

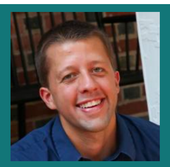
Example: Lozano v. Montoya Alvarez (2013)



Future Directions

1. Empirical estimates from real data
2. Address causal dependence between temporal utterances
3. Analyze between-judge and between-court temporal estimates

Katherine A. Keith
Then: University of Massachusetts Amherst, College of Information and Computer Sciences
Now: Postdoctoral researcher, AI2
Next: Assistant Professor, Williams College
kkeith@cs.umass.edu



Doug Rice
University of Massachusetts Amherst, Department of Political Science
drrice@legal.umass.edu

Brendan O'Connor
University of Massachusetts Amherst, College of Information and Computer Sciences
brenocon@cs.umass.edu

Paper & References:

Causal Inference + NLP Workshop, EMNLP 2021