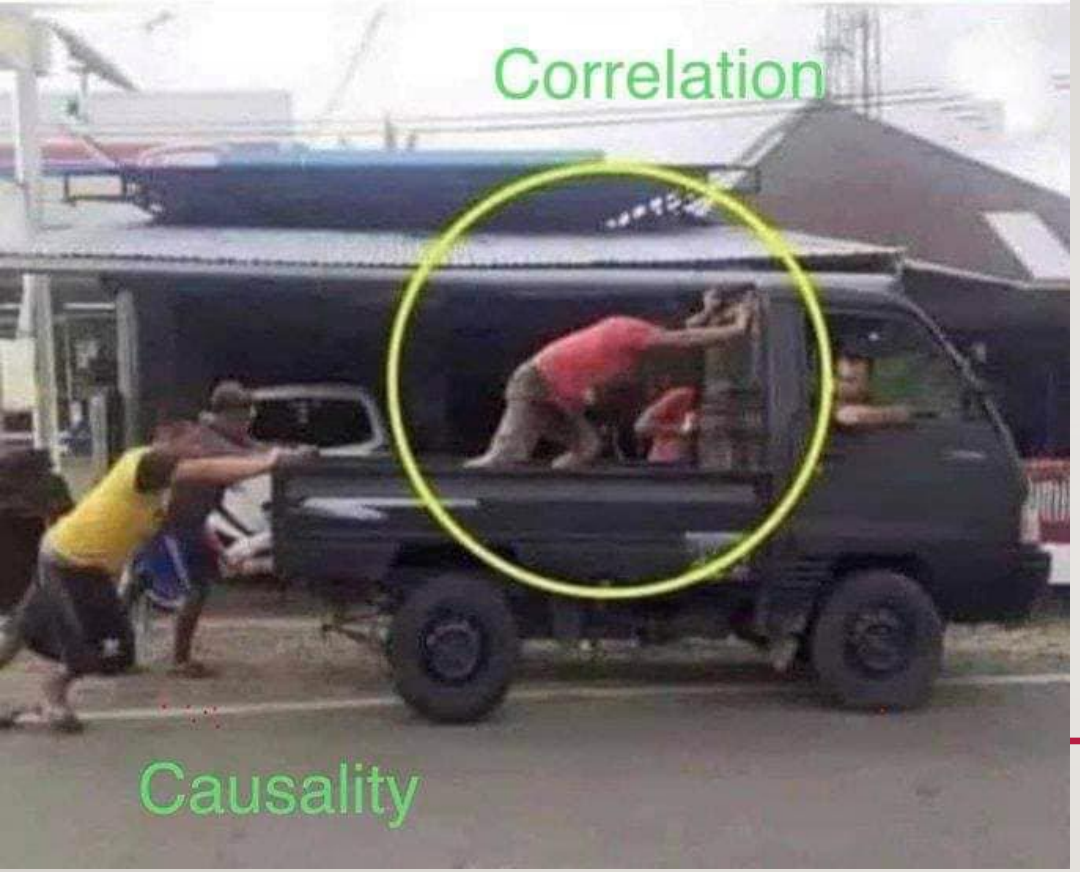


CAUSAL INFERENCE IN NATURAL LANGUAGE PROCESSING FOR ECONOMICS

WEEK 4: CAUSAL INFERENCE – TEXT AS MEDIATORS

SUMMER SEMESTER 22

DR.HUYEN NGUYEN

HTTP://HUYENTTNGUYEN.COM

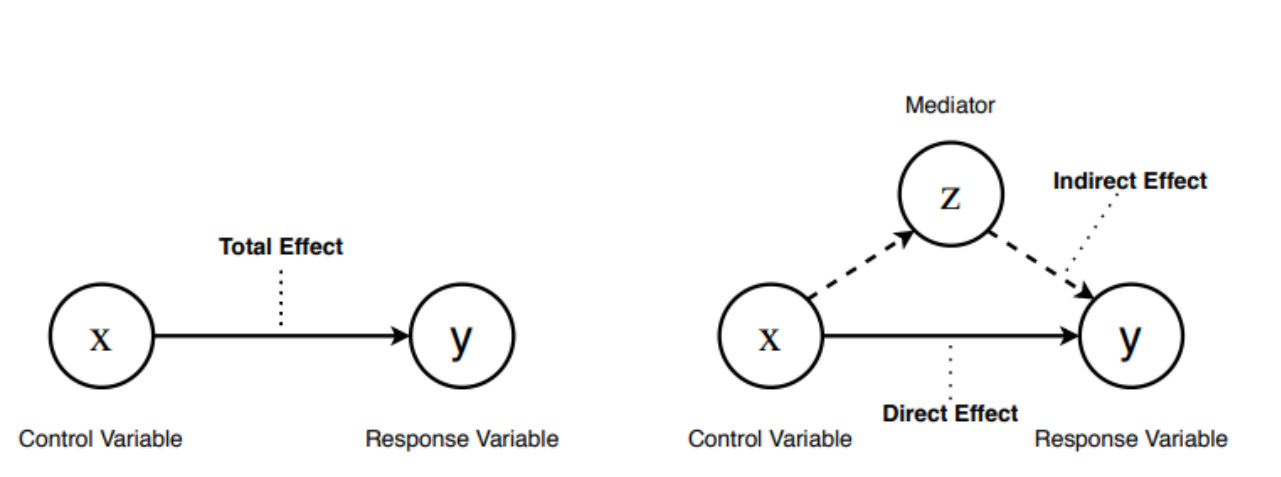
LOGISTICS

- Group allocation can be found on https://docs.google.com/spreadsheets/d/1EEUJp1Srz05BwEUH2u3CbYwMykmqEUhZdn2SqrmeaRs/edit#gid=0.
- Assigned papers (2-3 max) per group will be tentatively available latest by Thursday 28.04, for those who have registered their research interests.
- Should you have any papers you are interested in reading and presenting, please inform me via e-mail/Slack at least 3 weeks in advance of your presentation.

AGENDA

- Causal Mediation Analysis (CMA)
  - Framework
  - Assumptions & Identification Strategies
  - Applications:
    - (1) CMA and Biases in Language Models
    - (2) Text as Causal Mediators in Supreme Court Arguments
- Coding exercise: Aggregated Classification Pipeline from Start to Finish

CAUSAL MEDIATION ANALYSIS?

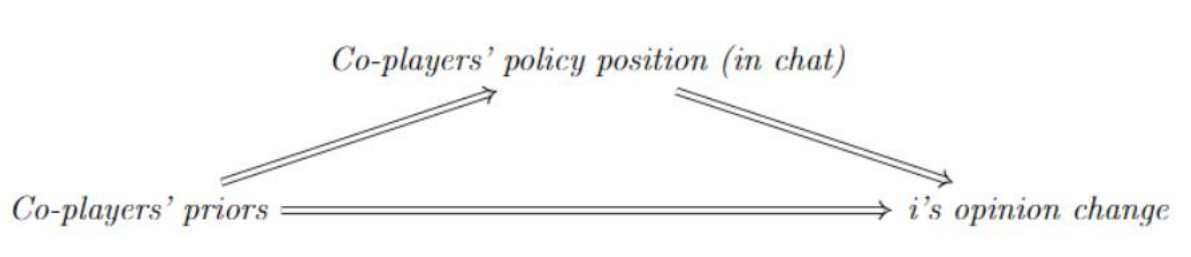


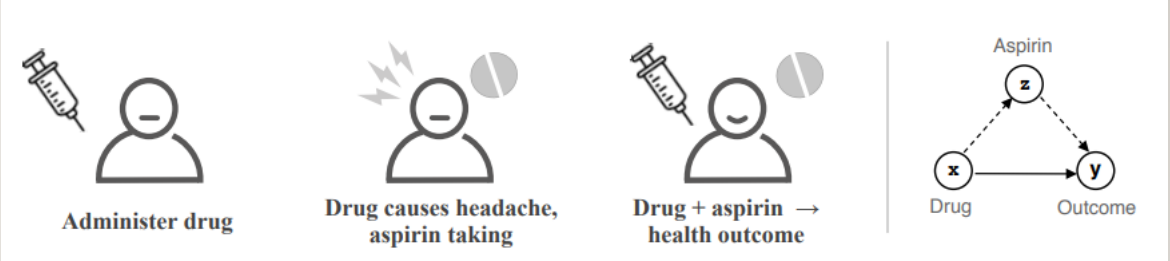
Causal mediation analysis (Pearl 2001, Huber 2020, Farbmacher et al., 2020) decomposes the causal effect of an intervention X on outcome variable y into:

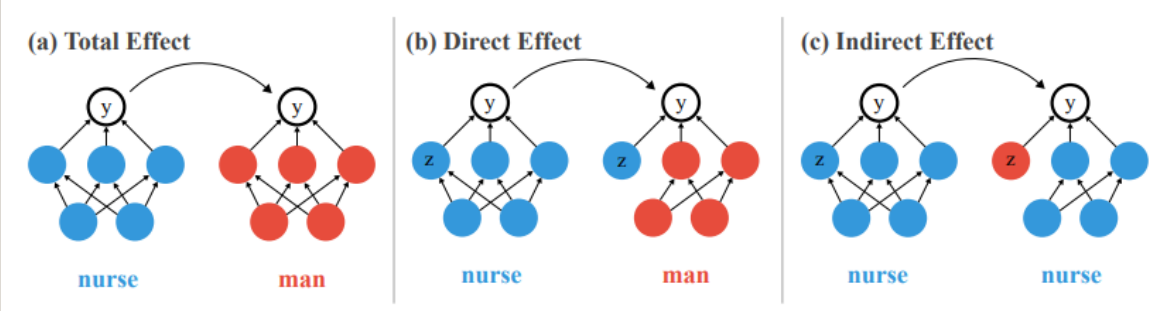
- The indirect effect operating through a mediator Z (or intermediate outcome).
- The direct effect of any causal mechanisms not operating through mediator Z.

→ reveal the causal role of **specific components** in a model's behavior.

EXAMPLES?


Co-players' policy position (in chat)
Co-players' priors
i's opinion change


Administer drug
Drug causes headache, aspirin taking
Drug + aspirin → health outcome
Aspirin
z
x
y
Drug
Outcome


(a) Total Effect
(b) Direct Effect
(c) Indirect Effect
y
z
nurse
man
nurse
nurse

ATE = DIRECT + INDIRECT EFFECTS?

GOAL: decompose the average treatment effect (ATE) of a binary treatment (1/0), denoted by D, on an outcome of interest Y, into an indirect effect operating through a discrete mediator M, and a direct effect that comprises any causal mechanisms other than through M.

$$\Delta = E[Y(1, M(1)) - Y(0, M(0))]$$

where $M(d) =$ potential mediator under the treatment value $d \in \{0,1\}$

and $Y(d, m) =$ potential outcome as a function of both the treatment and some value m of the mediator M.

DECOMPOSITION

AVERAGE DIRECT EFFECT $\theta(d)$

$$\theta(d) = E\left[Y\left(1, M(d)\right) - Y\left(0, M(d)\right)\right] \quad d \in \{0,1\}$$

= difference in mean potential outcomes when switching the treatment, while *keeping the potential mediator fixed* → blocks the causal mechanism via M.

DECOMPOSITION

AVERAGE DIRECT EFFECT $\theta(d)$

$$\theta(d) = E\big[Y\big(1, M(d)\big) - Y\big(0, M(d)\big)\big] \quad d \in \{0,1\}$$

= difference in mean potential outcomes when switching the treatment, while *keeping the potential mediator fixed* → blocks the causal mechanism via M.

AVERAGE INDIRECT EFFECT $\delta(d)$

$$\delta(d) = E\big[Y\big(d, M(1)\big) - Y\big(d, M(0)\big)\big] \; d \in \{0,1\}$$

= difference in mean potential outcomes when switching the potential mediator values, while keeping the treatment fixed → block the direct/natural effect.

DECOMPOSITION

AVERAGE DIRECT EFFECT $\theta(d)$

$$\theta(d) = E\big[Y\big(1, M(d)\big) - Y\big(0, M(d)\big)\big] \quad d \in \{0,1\}$$

= difference in mean potential outcomes when switching the treatment, while *keeping the potential mediator fixed* → blocks the causal mechanism via M.

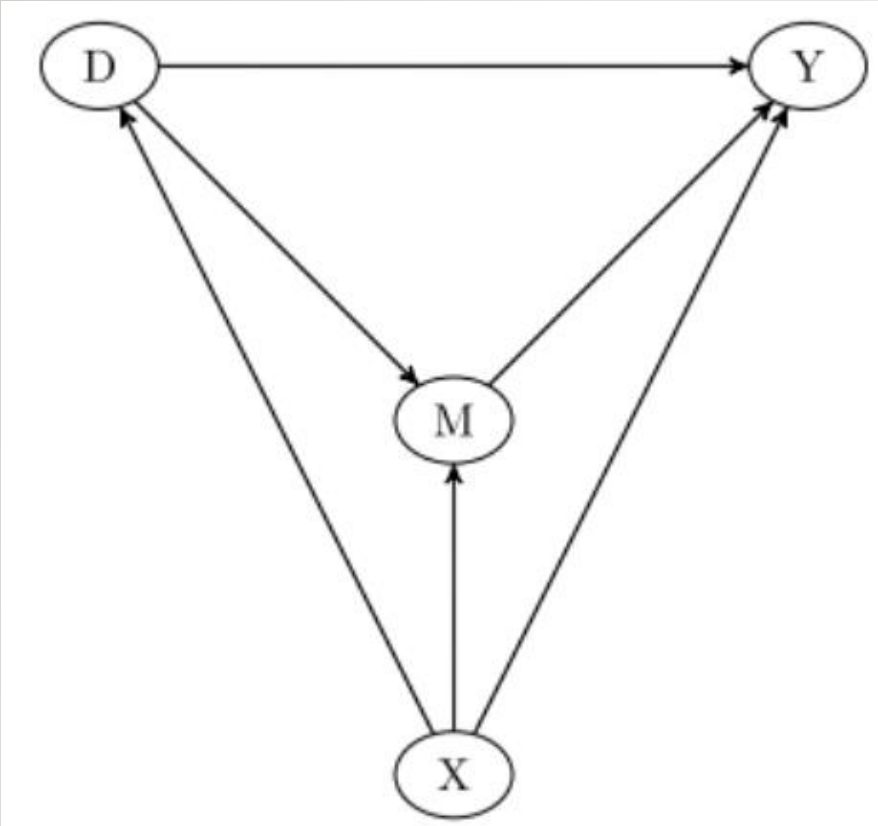
AVERAGE INDIRECT EFFECT $\delta(d)$

$$\delta(d) = E\big[Y\big(d, M(1)\big) - Y\big(d, M(0)\big)\big] \; d \in \{0,1\}$$

= difference in mean potential outcomes when switching the potential mediator values, while keeping the treatment fixed → block the direct/natural effect.

→ This distinction hints towards heterogeneous effects across treatment states d due to interaction between D and M.

ASSUMPTIONS & IDENTIFICATION STRATEGY



Key idea: Each of D, M and Y might be causally affected by distinct and statistically independent sets of unobservables not displayed in the Figure on the side,…

…BUT none of these unobservables may jointly affect two or all three elements (D, M, Y) conditional on X.

ASSUMPTIONS (1): CONDITIONAL INDEPENDENCE OF THE TREATMENT

$$\{Y(d', m), M(d)\} \perp (D, X) \forall\, d', d \in \{0,1\}$$

And m is in the support of M.

→Conditional on X, potential outcomes and mediators are independent of the treatment.

→No confounders jointly affecting the treatment on the one hand and the mediator and/or the outcome on the other hand

→In non-experimental data, the plausibility of Assumption 1 depends on the richness of X.

ASSUMPTIONS (2): CONDITIONAL INDEPENDENCE OF THE MEDIATORS

$$Y(d', m) \perp M | D = d, X = x \; \forall \; d', d \in \{0,1\}$$

and m, x in the support of M, X.

→ Conditional on treatment D and X, no confounders can jointly affect the mediator and the outcome.

→ (!) No post-treatment confounders of the mediator M – Outcome Y relations i.e. can be challenging if the time window between the measurement of the treatment and the mediator is large in a world of time-varying variables.

ASSUMPTION (3): COMMON SUPPORT (A.K.A POSITIVITY ASSUMPTION)

$$\Pr(D = d|M = m, X = x) > 0 \; \forall d \in \{0,1\}$$

And m, x in the support of M, X.

→The conditional probability to be/not be treated given M, X is larger than zero.

→The treatment must NOT be deterministic in X, otherwise no comparable units in terms of X are available across treatment states.

→$\Pr(D = d|M = m, X = x) > 0$ if M is discrete OR the conditional density of M given D, X is larger than 0 if M is continuous.

→Conditional on X, the mediator state must NOT be deterministic in the treatment, otherwise no comparable units in terms of the treatment are available across mediator states.

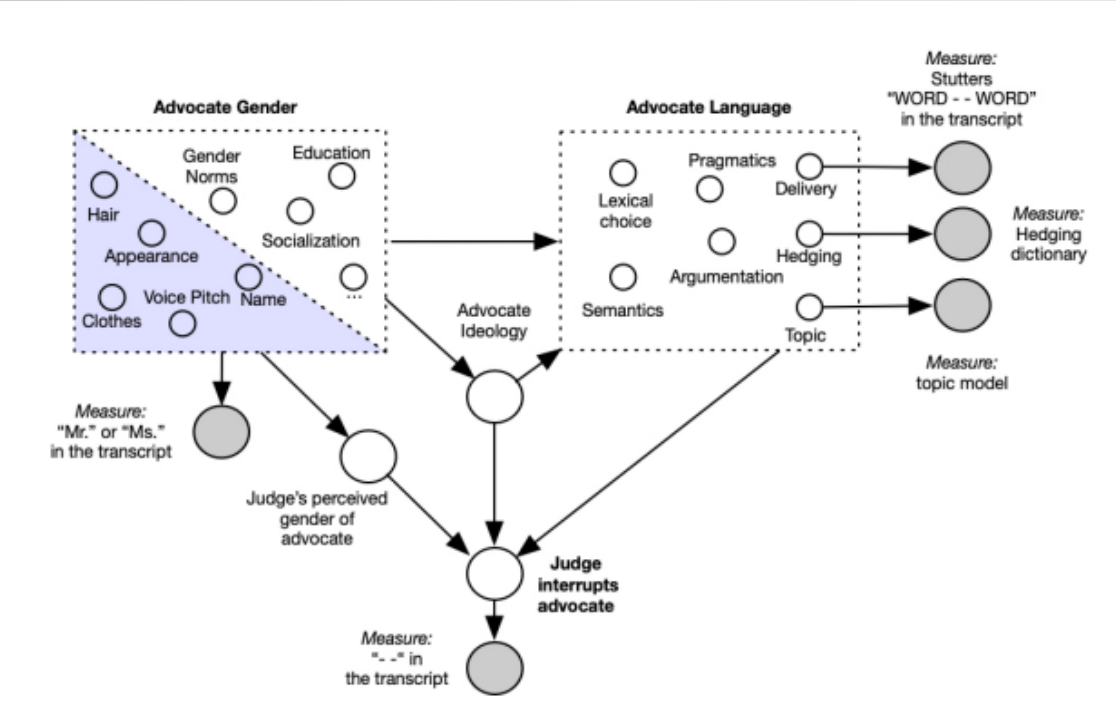
CMA + TEXT = BIASES IN LANGUAGE MODEL (VIG ET AL., NEURNIPS 2020)

- Given a prompt u such as *The nurse said* that, a language model (LM) generates a continuation.
- A biased LM may assign a higher likelihood to *she* than to *he*, such that $p_\theta(she|u) > p_\theta(he|u)$.
- In this case, she is the stereotypical candidate, while he is the anti-stereotypical candidate, which reflects a societal bias associating nurses with women more than men.

TEXT AS CAUSAL MEDIATORS (KEITH, RICE AND O'CONNOR 2021) IN OBSERVATIONAL DATA

- Important decision makers sometimes treat some social groups (e.g. women, racial minorities, or ideological communities) differently than others.
- Case in Point: During Supreme Court oral arguments, is a justice interrupting female advocates more because of their gender? Content of the advocates' legal arguments? Advocates' language delivery?

TEXT AS CAUSAL MEDIATORS (KEITH, RICE AND O'CONNOR 2021) IN OBSERVATIONAL DATA



Possible language mediators M^j in this context:

(A) Hedging

(B) Speech disfluencies

(C) Semantic Topics

Three common concerns:

- Is it interpretable?
- To what extent does one expect measurement errors?
- Is it causally independent from other measured language aspects?

EXERCISE TIME ☺

- Open your Google Colab.
- Go to Github's course channel →Exercise folder to access these files.