# Counterfactually Fair Prediction Using Multiple Causal Models

Fabio Massimo Zennaro[1]
Magdalena Ivanovska[2]

*University of Oslo*

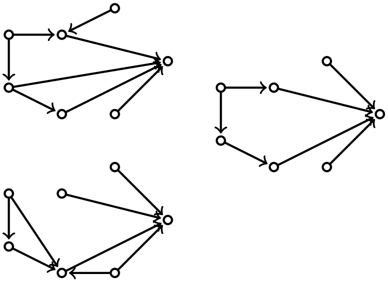
December 7, 2018

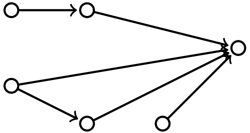
[1]fabiomz@ifi.uio.no
[2]magdalei@ifi.uio.no

# 1. Definition of the problem

# Statement of the problem

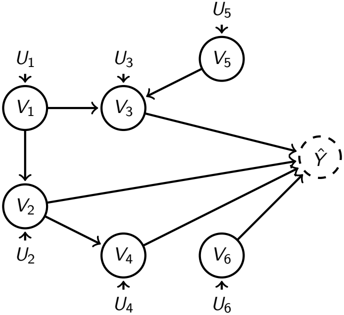
Given $N$ agents defining **causal models** for *prediction*, how can we aggregate them in a single model that is guaranteed to be **fair**?





# Probabilistic Structural Causal Model (Pearl [2009])

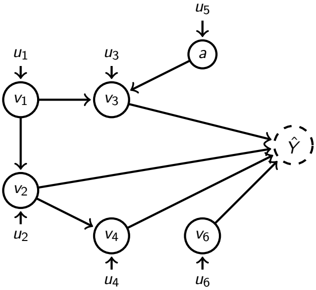
$$\mathcal{M} = (\mathcal{U}, \mathcal{V}, \mathcal{F}, P(U))$$

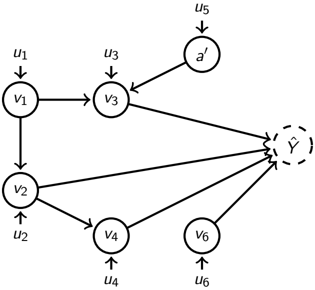



- Encoding *causal relationships*;
- Deterministic *endogenous nodes* and stochastic *exogenous nodes*;
- Allows the definition of *interventions* and *counterfactuals*.

# Countefactual Fairness (Kusner et al. [2017])

$$P\left(\hat{Y}_{A\leftarrow a}(u)|V=v\right)=P\left(\hat{Y}_{A\leftarrow a'}(u)|V=v\right)$$




- Probability of the predictive output ($\hat{Y}$) when we *intervene* to change a *sensitive attribute* ($a \rightarrow a'$), provided that all the other endogenous ($v$) and exogenous ($u$) variables remain the same.

# Definition of the problem

Given $N$ agents defining predictive **probabilistic structural causal models** such that:

- they work on the same set of exogenous and endogenous variables;
- they are not aware of fairness requirement;

how does a centralized authority assemble these models in a single model that is **counterfactually fair** with respect to a set of chosen *sensitive attributes*?

# 2. Proposed solution

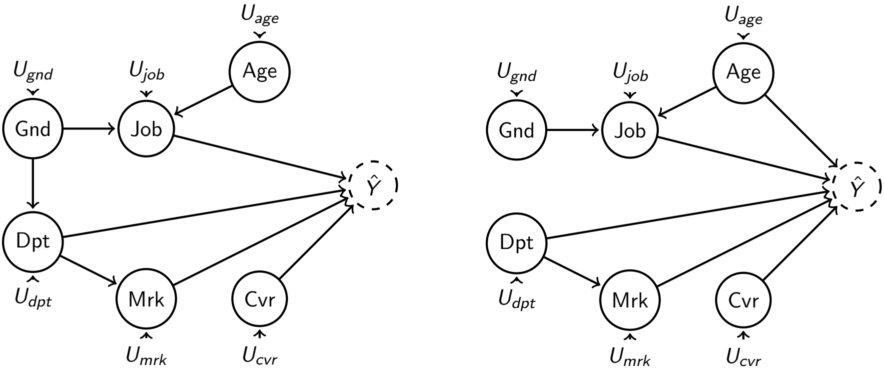
# Overview of the solution

We suggest a solution based on a *two-stage approach* (Bradley et al. [2014])

1. **Qualitative stage**: defining an aggregated core counterfactually-fair graph;
   A. *Pooling step*: performing *judgment aggregation* over edges;
   B. *Removal step*: enforcing *counterfactual fairness*.
2. **Quantitative stage**: predicting a counterfactually-fair output;
   A. *Sampling step*: performing *Monte Carlo sampling* from marginalized graphs;
   B. *Pooling step*: performing *opinion pooling* of the sampled outputs.

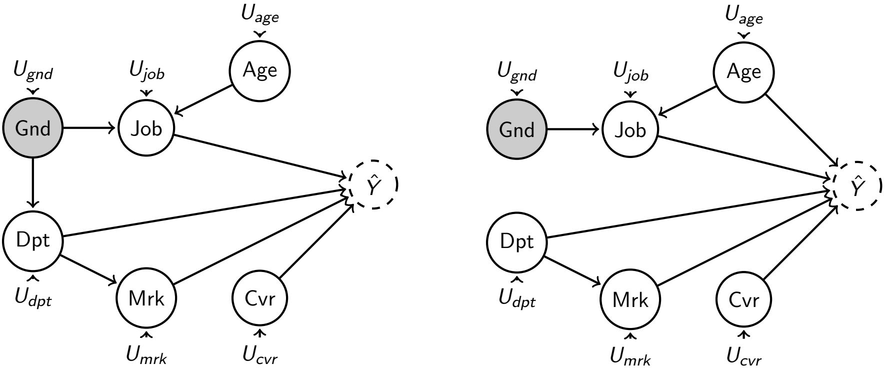
# Toy Example: Agents

Agents define *models*:



# Toy Example: Decision Maker

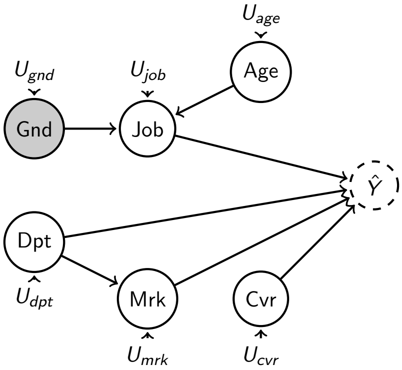
Decision maker chooses *protected attributes*:




And sets a judgment aggregation rule (*majority rule*) and an opinion aggregation rule (*averaging rule*).

# 1A. Pooling step

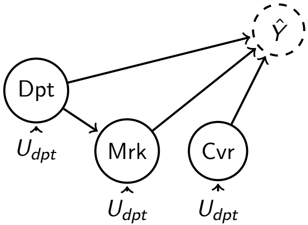
Given a judgment aggregation rule, perform aggregation over the edges ordered wrt to their distance from the predictor $\hat{Y}$[3].




[3]Ordering is a necessary technical condition to counter the *judgment aggregation impossiblity theorem* (Bradley et al. [2014])

# 1B. Removal step

Remove protected attributes and their descendants.[4].

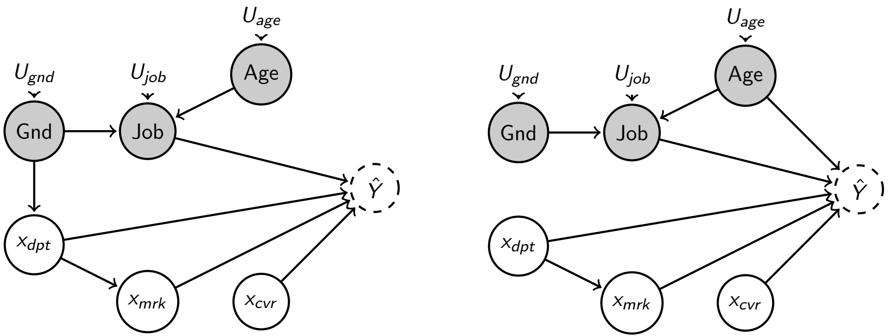



[4]Removing these nodes is a technical condition to guarantee *counterfactual fairness* (Kusner et al. [2017])

# 2A. Sampling step

Given an input $X$, compute the predictive output $\hat{Y}$ randomly sampling all the nodes that do not belong the fair graph:

$$P(\hat{Y}_i|X) = \int P(\hat{Y}_i|X_f = x_f, X_{\bar{f}})dX_{\bar{f}}$$



# 2B. Pooling step

Given an opinion aggregation rule, perform aggregation over the predictive probability distribution of each one of the N agents:

$$\hat{Y} = \frac{1}{N}\sum_{i=1}^{N} E\left[P(\hat{Y}_i|X)\right]$$

The output is guaranteed to be counterfactually fair[5].

[5] See the paper for a complete illustration over the toy case

# Conclusions

Preliminary work with several avenues of development:

- Can we preserve more information in the removal step?
- Can we extend the approach to agents defining models over different variables?
- Can we consider distributed scenarios?
- Can we relax the fairness constraint?
- Can we integrate observational fairness with affirmative fairness?

# Thanks!

Thank you for listening!

# References I

Richard Bradley, Franz Dietrich, and Christian List. Aggregating causal judgments. *Philosophy of Science*, 81(4):491–515, 2014.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4069–4079, 2017.

Judea Pearl. *Causality*. Cambridge university press, 2009.