

Ethical AI and Actuarial Practice

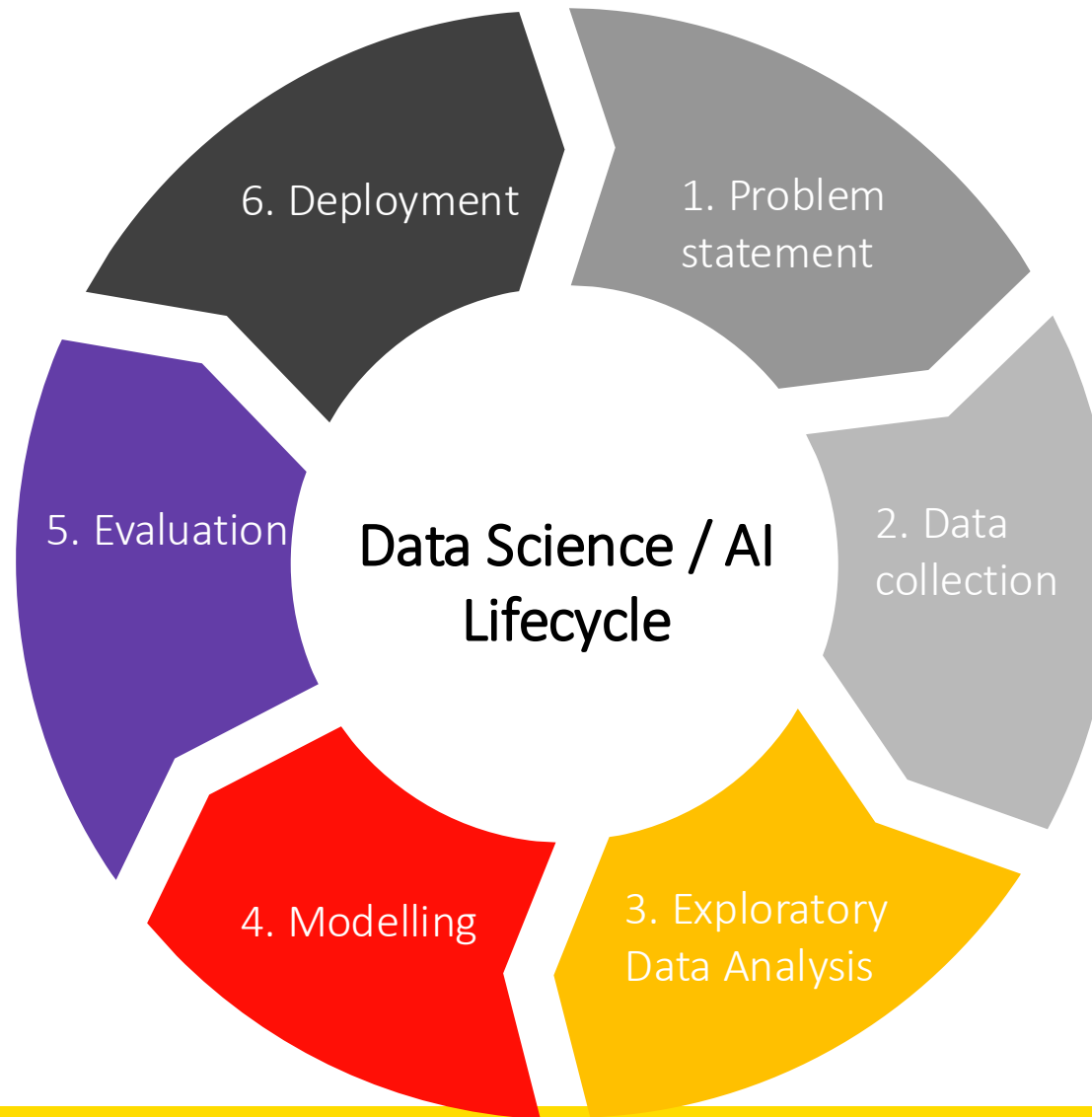
Fei Huang



UNSW
SYDNEY

Q: What is the most frequent failure in data analysis?

A: Mistaking the type of problems being considered



Actuarial Control Cycle:

- Stage 1 corresponds to defining the problem,
- Stages 2-5 to designing the solution
- Stage 6 to monitoring the results – mirroring the actuarial control cycle structure.

What are the problem types?

1. **Descriptive** — What happened?
Summarizes the main features of a dataset without making interpretations or drawing conclusions.
Example: What is the average temperature in July across Australian cities?
2. **Exploratory** — What patterns or relationships exist in the data?
Identifies trends, clusters, correlations, or anomalies to generate hypotheses or insights.
Example: Are there clusters of customer behavior in purchase data?
3. **Inferential** — What can we infer about a population based on a sample?
Uses statistical techniques to draw conclusions about a broader population from a sample.
Example: What is the average income of all Australian adults based on survey data?

What are the question types?

4. **Predictive** — What is likely to happen in the future?
Uses historical data and models to forecast future outcomes or behaviors.
Example: What is the expected number of insurance claims for a driver next year, based on their age, location, and driving history?
5. **Causal** — What is the effect of X on Y?
Determines whether and how a change in one variable causes a change in another.
Example: Does wearing a seatbelt reduce fatalities in car crashes?
6. **Mechanistic** — How does the system work exactly?
Explains the precise, deterministic relationship between variables based on underlying theory or physical laws. Rare outside fields like physics or engineering.
Example: How does wing design change airflow to reduce drag?

Major principles of AI ethics

1. **Fairness and non-discrimination:** mitigating any direct and indirect discrimination that is embedded in data or model design
2. **Transparency and explainability:** decisions interpretable by not only developers but also customers and regulators
3. **Accountability:** who is responsible when AI makes errors: actuary, developer or firm?
4. **Privacy and data ethics:** related to third-party sources, telematics or even social media
5. **Contestability:** people affected by automated decisions have meaningful channels for appeal and recourse
6. **Stability and robustness:** AI models remain reliable and perform as expected under real-world conditions.

Ethical AI Lifecycle

1. AI Problem Definition Considerations

- Align objectives with fairness, explainability, and regulations
- Define optimisation trade-offs (e.g., profit vs. societal equity)
- Assign stakeholder roles for AI use, oversight, and accountability
- Establish clear risk tolerances and success criteria

1. Problem statement

2. Ethical Data Collection Principles

- Minimize data privacy invasiveness
- Avoid indirect encoding of sensitive attributes (e.g., via proxies)
- Ensure meaningful consent for data use
- Identify and account for historical biases
- Maintain data quality for fair and robust modeling
- Align data use with governance and legal frameworks

2. Data collection

3. Ethical Feature Assessment in EDA

- Document assumptions, transformations, and selection criteria transparently
- Justify inclusion/exclusion of features with fairness or legal risk implications
- Ensure feature engineering decisions support model interpretability and auditability

3. Exploratory Data Analysis

4. Modelling

4. Ethical AI Modelling

- Mitigate historical or structural bias and achieve fairness
- Ensure interpretability for transparency, auditability, and regulation
- Justify key modelling choices and conduct independent validation
- Transparently document trade-offs between performance and ethical considerations

5. Evaluation

5. Ethical AI Model Evaluation

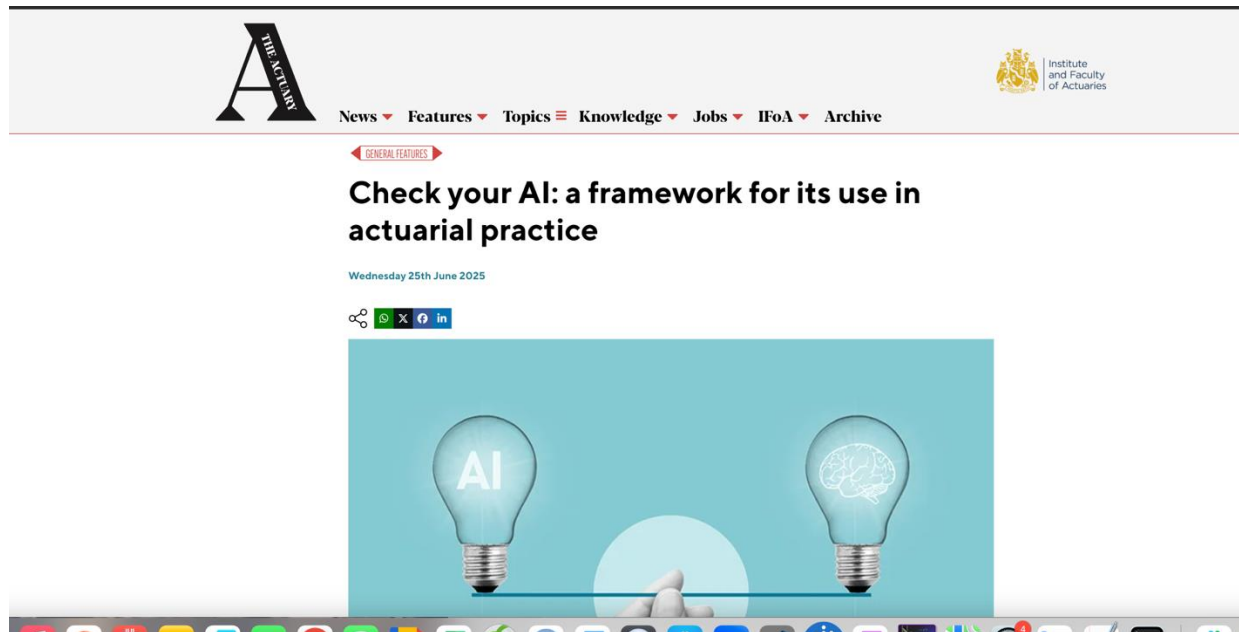
- Assess generalisation across diverse scenarios and populations
- Conduct scenario tests and sensitivity analyses for model resilience
- Validate model performance against ethical principles using holdout samples or real-world simulation
- Ensure transparency of evaluation metrics and assumptions

6. Deployment

6. Ethical Deployment

- Audit models for accuracy, fairness, transparency, and robustness
- Enable contestability for stakeholders
- Align deployment decisions with original problem definition and ethical goals
- Monitor model behavior post-deployment and incorporate feedback into future cycles
- Document deployment assumptions, risks, and governance protocols

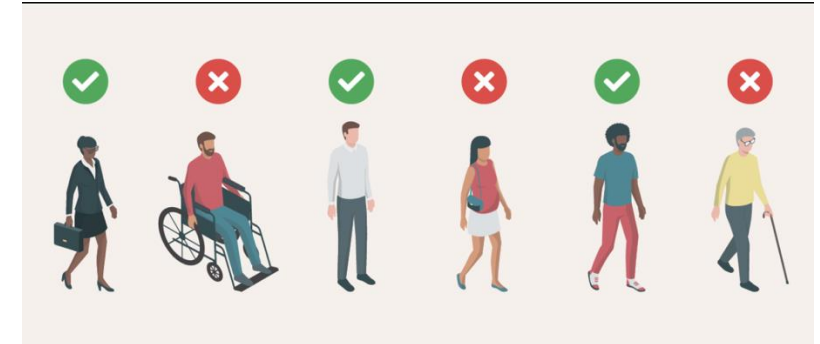
“The actuarial profession should lead, not follow, in building public trust in AI systems.”









Q1: What are the **principles** to assess the appropriateness of insurance discrimination (differentiation)?

Background

- Discrimination is rampant in society.
- Insurance pricing is interesting because the entire industry is based on discrimination.
- Although many variables are routinely used, the use of others (e.g., ethnicity/heritage/religion) are forbidden.
- How do insurers discriminate among customers?
 - Issuance, renewal, or cancellation
 - Coverage
 - Underwriting/Pricing
 - Marketing
 - Claims processing



MEN vs WOMEN		
Who Pays More for Insurance?		
TYPE OF INSURANCE	WHO PAYS MORE?	WHY ARE WE PAYING MORE?
 Life Insurance		Shorter life expectancy
 Car Insurance		More likely to get into accidents
 Health Insurance		Longer life expectancy

Fairness depends on the context.

Is Insurance a **Social Good** or an **Economic Commodity**?

Is risk pooling based on **Subsidizing Solidarity** or **Chance Solidarity**?

Insurance as a Social Good?

- Benefits the general public, such as clean air and clean water
- Non-excludable
- Mandatory insurance (social safety net)
- Limited risk classification (or community rating)

- ❖ Health insurance is likely to be seen as a social good.
- ❖ Compulsory third-party auto insurance is a social good.
- ❖ Catastrophe insurance is treated as a social good by some jurisdiction.

Insurance as an Economic Commodity?

- Voluntary insurance (luxury)
- Market price depends on the forces of supply and demand
- “Free” market (little regulation)
- Risk classification advocated for reasons related to adverse selection, moral hazard, and economic efficiency.

- ❖ Life insurance is more often seen as a private (non-public) product, an economic commodity.
- ❖ Voluntary auto insurance is an economic commodity.
- ❖ Long-term care and disability insurance falls in between.

Is it fair to use that rating factor?

- **Control.** e.g., your ownership of a sports car; many sensitive attributes are uncontrollable (gender, race, ethnicity, nationality, etc.)
- **Mutability.** Does the variable change over time (such as age) or stay fixed?
- **Causality.** It is generally acceptable to use a variable if it is known to cause an insured event (e.g., cancer in life insurance).
- **Statistical Discrimination.** A variable must have some predictive value of an underlying risk. A necessary, but not sufficient, condition.
- **Limiting or Reversing the Effects of Past Discrimination.** e.g., discriminating based on skin color is more problematic than based on eye color.
- **Inhibiting Socially Valuable Behavior.** e.g. whether to participate in genetic testing research when insurers make decisions based on genetic test results.

Indirect Discrimination

-- a grey area in regulation

- Direct discrimination is prohibited, but indirect discrimination using proxies or more complex and opaque algorithms are not clearly regulated or assessed.
- Do we have a clear definition of indirect discrimination?
- How to assess and how to mitigate indirect discrimination?

How to balance the social justice and economic efficiency to mitigate discrimination?

- **Input-based approach:** restricting the use of certain rating factors or pricing procedures
 - protected or proxy variables (*fairness through unawareness*)
 - non-risk-based price optimization/price walking
- **Output-based approach:** quantitative fairness criteria

Q2: How to **evaluate** fairness and **mitigate indirect discrimination** in the context of **AI** and Big Data?

Existing Regulations

- ▶ No Regulation
- ▶ Restriction on the Use of a Protected Variable
- ▶ Prohibition on the Use of a Protected Variable
- ▶ Restriction on the Use of a Proxy Variable
- ▶ Prohibition on the Use of a Proxy Variable
- ▶ Disparate Impact Standard
- ▶ Community Rating
- ▶ Affirmative Action

Quantitative Fairness Criteria

- Association-based
 - Individual fairness
 - Group fairness
 - Independence
 - Separation
 - Sufficiency (Calibration)
- Causality-based
 - Counter-factual fairness

- Are they suitable for the insurance context?
- Are they compatible with each other?
- Which one to use?

Araiza Iturria et al. (2022), Baumann and Loi (2023), Charpentier, A. (2023), Côté et al. (2024), Lindholm et al. (2022, 2023), Xin and Huang (2023), CAS Research Paper Series, SOA Research Paper Series

Fairness Criteria for Insurance Pricing

- ▶ Let X_P denote the protected attribute, which is a categorical variable and has only two groups $X_P = \{a, b\}$.
- ▶ Let X_{NP} denote other available (non-protected) attributes, and hence the feature space is $X = \{X_P, X_{NP}\}$.
- ▶ Let \hat{Y} denote the predictor or the decision outcome of interest, $\hat{Y} \in \mathbb{R}$. In our context, \hat{Y} is the premium charged by the insurer, and in this paper, we assume that \hat{Y} is approximately equal to the pure premium and ignore any expenses or profit loadings.
- ▶ Let Y denote the observed outcome of interest, $Y \in \mathbb{R}$. Note that Y is not known when the policy is issued, Y is a measure of real claim experience observed by the insurer over a given period after policy issuance.

Fairness Criteria – Individual Fairness

Definition 1. Fairness through Unawareness (FTU): Fairness is achieved if the protected attribute X_P is not explicitly used in calculating the insurance premium \hat{Y} .

Definition 2. Fairness through Awareness (FTA): A predictor \hat{Y} satisfies fairness through awareness if it gives similar predictions to similar individuals (Dwork et al. 2012; Kusner et al. 2017).

Definition 3. Counterfactual Fairness (CF): A predictor \hat{Y} is counterfactually fair for an individual if “its prediction in the real world is the same as that in the counterfactual world where the individual had belonged to a different demographic group” (Kusner et al. 2017; Wu, Zhang, and Wu 2019, p. 1) or, mathematically, given that $X = x$ and $X_P = a$, for all y and for simplicity, X_P has only two groups $\{a, b\}$, a predictor \hat{Y} is counterfactually fair if

$$\mathbb{P}(\hat{Y}_{X_P \leftarrow b}(U) = y \mid X_{NP} = x, X_P = b) = \mathbb{P}(\hat{Y}_{X_P \leftarrow a}(U) = y \mid X_{NP} = x, X_P = b).$$

Definition 4. Controlling for the Protected Variable (CPV): As defined in definition 6 in Lindholm et al. (2022a), a *discrimination-free price* for Y with respect to X_{NP} is defined by

$$h^*(X_{NP}) := \int_{x_P} \mathbb{E}[Y \mid X_{NP}, x_P] d\mathbb{P}^*(x_P),$$

where $\mathbb{P}^*(x_P)$ is defined on the same range as $\mathbb{P}(x_P)$.

Fairness Criteria – Group Fairness

Definition 5. Demographic Parity (DP): A predictor \hat{Y} satisfies demographic parity if

$$\mathbb{P}(\hat{Y}|X_P = a) = \mathbb{P}(\hat{Y}|X_P = b).$$

Definition 6. Relaxed Demographic Parity (RDP): A predictor \hat{Y} satisfies relaxed demographic parity or has no disparate impact if the following ratio is above certain threshold τ (Feldman et al. 2015):

$$\frac{\mathbb{P}(\hat{Y} = \hat{y}|X_P = b)}{\mathbb{P}(\hat{Y} = \hat{y}|X_P = a)} > \tau.$$

Definition 7. Conditional Demographic Parity (CDP): A predictor \hat{Y} satisfies conditional demographic parity if

$$\mathbb{P}(\hat{Y}|X_{NP_{legit}} = x_{NP_{legit}}, X_P = a) = \mathbb{P}(\hat{Y}|X_{NP_{legit}} = x_{NP_{legit}}, X_P = b),$$

where $X_{NP_{legit}}$ denotes a subset of “legitimate” attributes within unprotected attributes in the feature space ($X_{NP_{legit}} \subseteq X_{NP} \subset X$) that are permitted to affect the outcome of interest (Corbett-Davies et al. 2017; Verma and Rubin 2018).

Regulations, Fairness Criteria, and Models

TABLE 1
Comparison between Different Regulations

Regulation	Fairness criteria	Representative model
No regulation	Neither	M0
Restriction on a protected variable	Neither	M0*
Prohibition on a protected variable	Individual	MU or MC
Restriction on a proxy variable	Individual	MU*
Prohibition on a proxy variable	Individual	MU*
Disparate impact standard	Group	MDP or MCDP
Community rating	Group	MDP or MCDP
Affirmative action	Neither	None

Questions to Tackle



Which fairness notion to use?

- Proxy-discrimination (omitted variable bias)
- Conditional demographic parity
- Equalized odds (error rates parity)
- Well-calibration



How to achieve it?

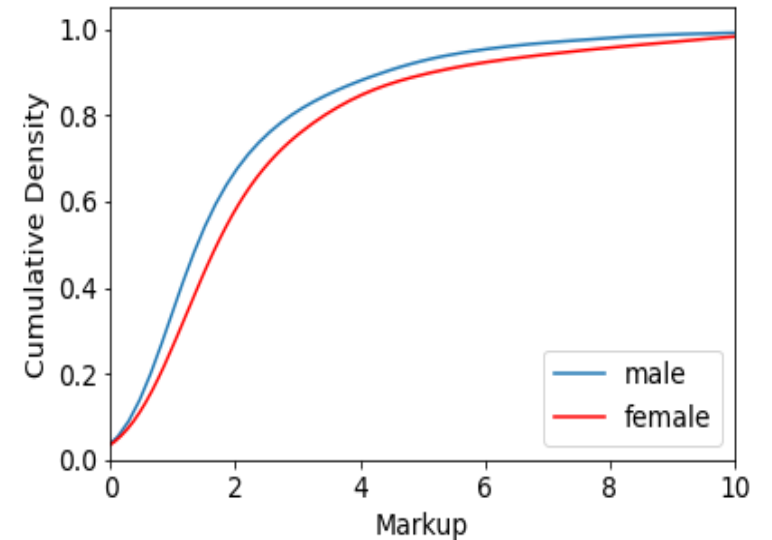
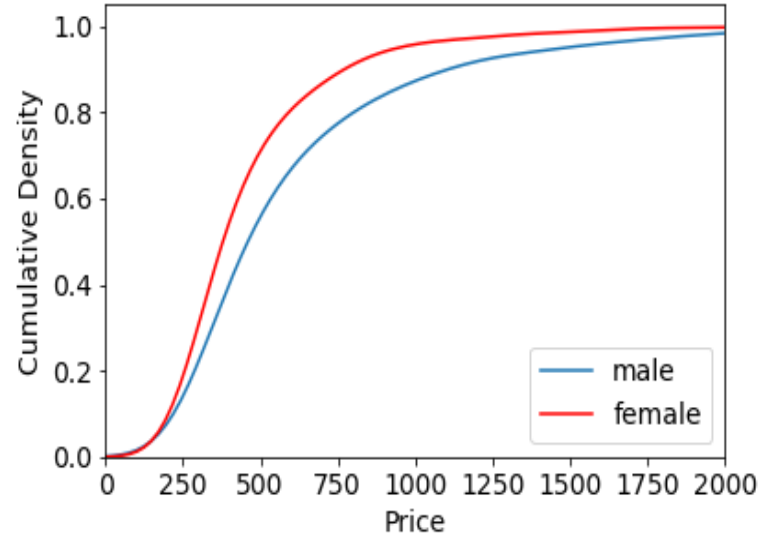
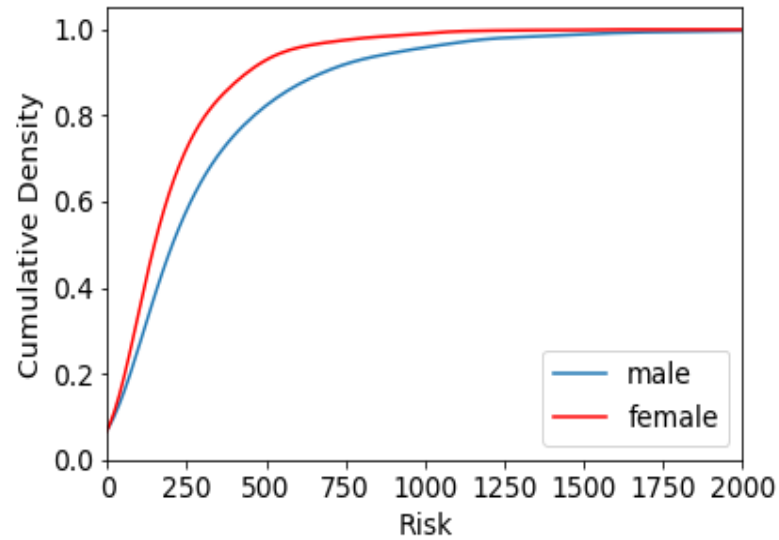
- Fairness in cost prediction or market pricing?
- Fairness in terms of prices or markups?
- What if we cannot observe protected variables?



What is the impact of fairness on stakeholders?

- Consumer welfare v.s. Firm profit
- Welfare for males v.s. Welfare for females

Discrepancy between Male and Female



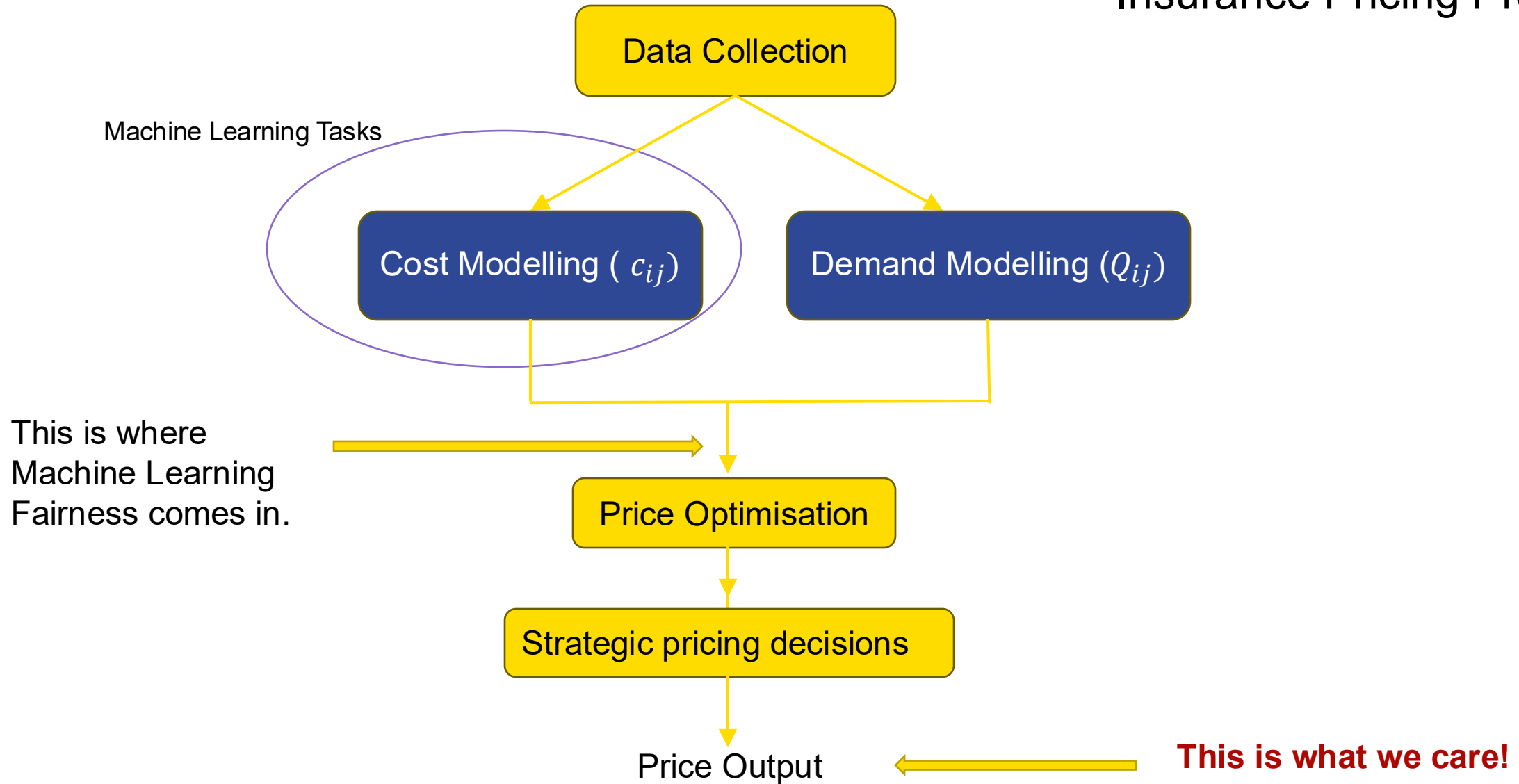
- Female consumers are:
 - safer drivers
 - charged less price
 - charged more markup

Which group is the "disadvantaged"?

Q3:

What is the **welfare impact** of fair policies on stakeholders?

Insurance Pricing Process



Is fair insurance pricing a fair machine learning problem?

Given the features X and the target variable Y ,

$$\begin{aligned} \min L(Y, f(X)) \\ \text{s.t. Fairness Const.} \\ \text{Accountability Const.} \end{aligned}$$

OR

$$\min L(Y, f(X)) + \lambda \text{Penalty}(f)$$



Fairness in ML \neq
Fairness in outcome

Cost Modeling

- Claim frequency prediction
- Claim severity prediction

Demand modeling

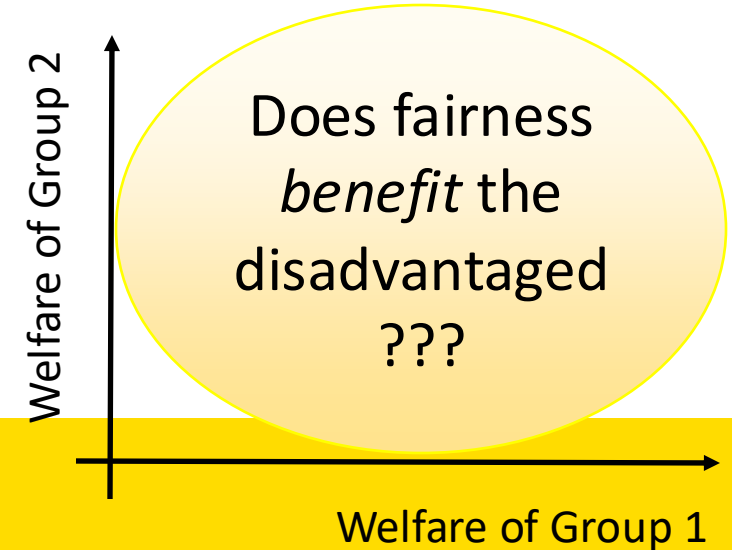
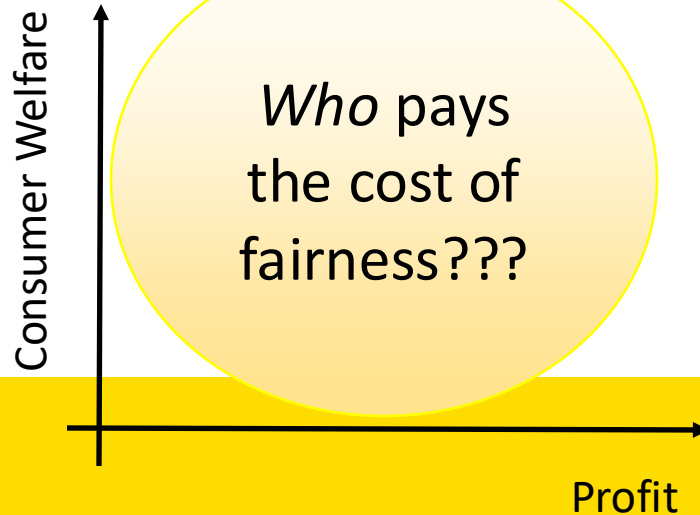
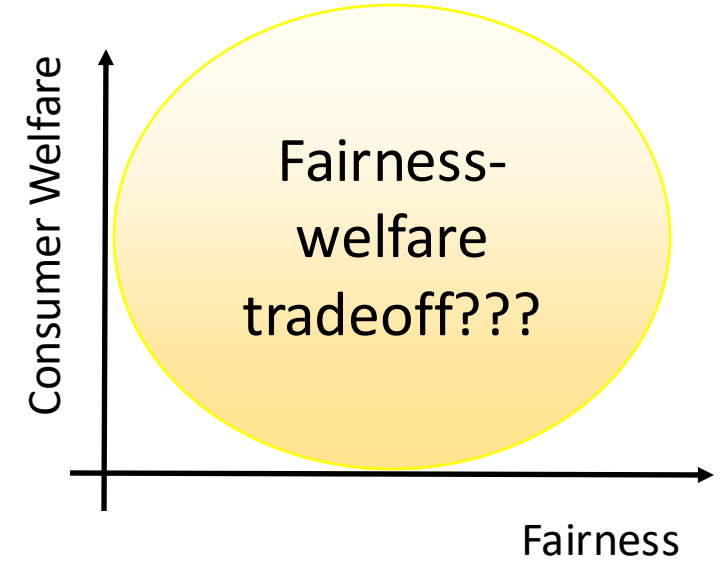
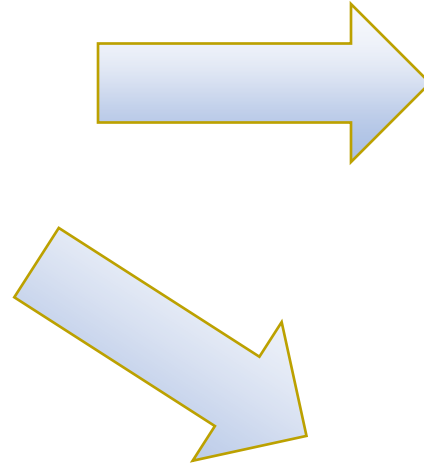
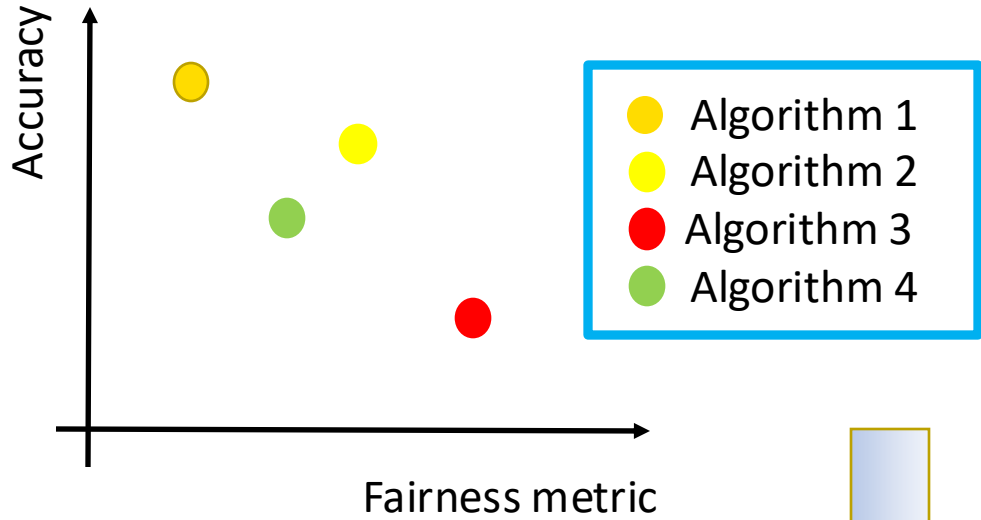
- Demand function estimation
- Price elasticity

Pricing decision

- Profit maximization
- Financial and other constraints

Impact of Fairness Policies on Stakeholders

Fairness-accuracy tradeoff



Is fair insurance pricing a fair machine learning problem?

Given the features X and the target variable Y ,

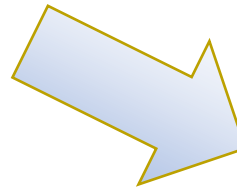
$$\begin{aligned} \min L(Y, f(X)) \\ \text{s.t. Fairness Const.} \\ \text{Accountability Const.} \end{aligned}$$

OR

$$\min L(Y, f(X)) + \lambda \text{Penalty}(f)$$



Fairness in ML \neq
Fairness in outcome



Welfare impact
of algorithms

Cost Modeling

- Claim frequency prediction
- Claim severity prediction

Demand modeling

- Demand function estimation
- Price elasticity

Pricing decision

- Profit maximization
- Financial and other constraints

Consumers' choice

- Product choice
- Purchasing decision

Consumers' welfare

Profit

Empirical Approach to Assess Welfare Impact

1) Model insurers' decision problem

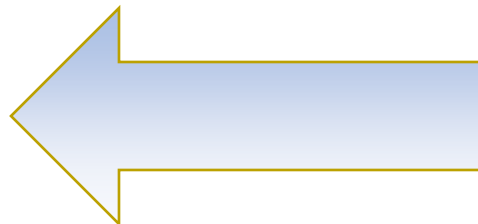
- Cost modeling
- Demand modeling
- Objective/constraint in pricing

2) Estimate consumers' choice model

- Specify utility function
- Estimate the utility parameters

3) Counterfactual simulation

- Simulate the price under different policy using (1)
- Evaluate consumer surplus using (2)



A) Cost model regulations

- Fairness criteria
- Accountability constraint

B) Pricing practice regulations

- Fairness criteria
- Accountability constraint
- Price optimization ban

Data and Implementation

- **Cost Modeling**

- Data: A French private motor insurance drawn from the R package CASdatasets (Charpentier and Dutang, 2015). We focus on the material damage coverage. It contains 100,000 third-party liability (TPL) policies observed from 2009 to 2010.
- Protected attribute: gender
- Model: GLM and XGBoost (Poisson & gamma loss)

- **Demand Modeling**

- We construct our simulated consumers by utilizing the claims data from Dutang et al. (2015) and the estimated demand models from Einav et al. (2010) and Jin and Vasserman (2021).

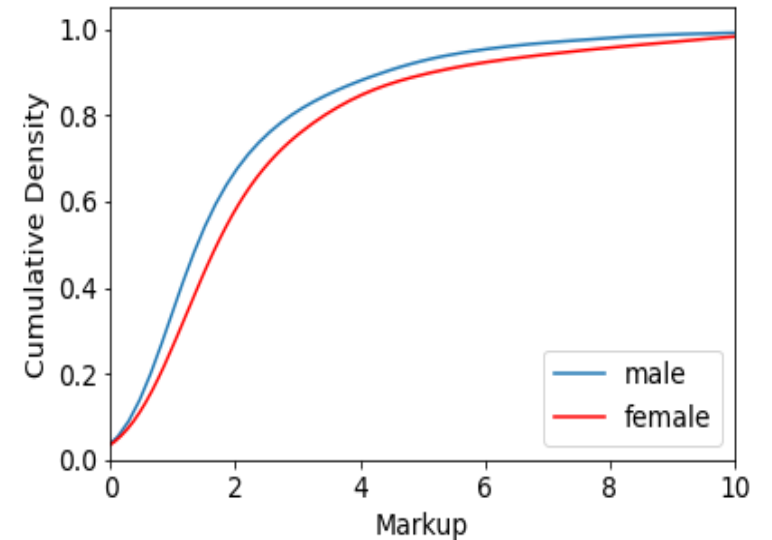
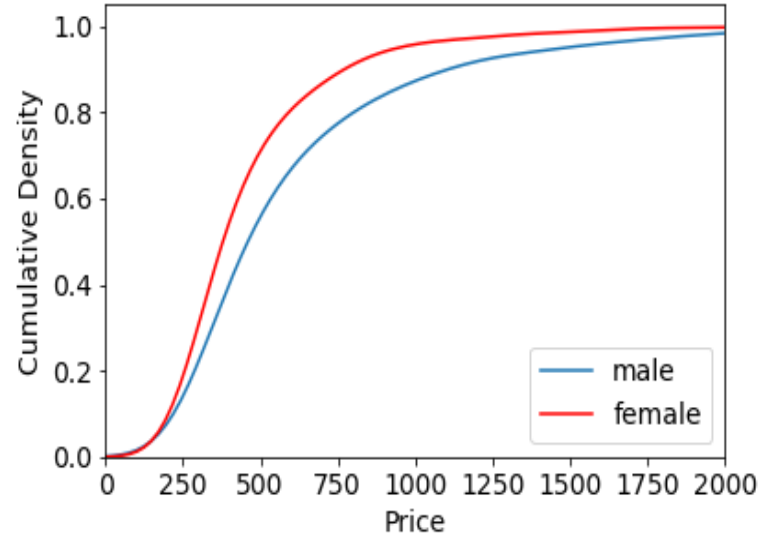
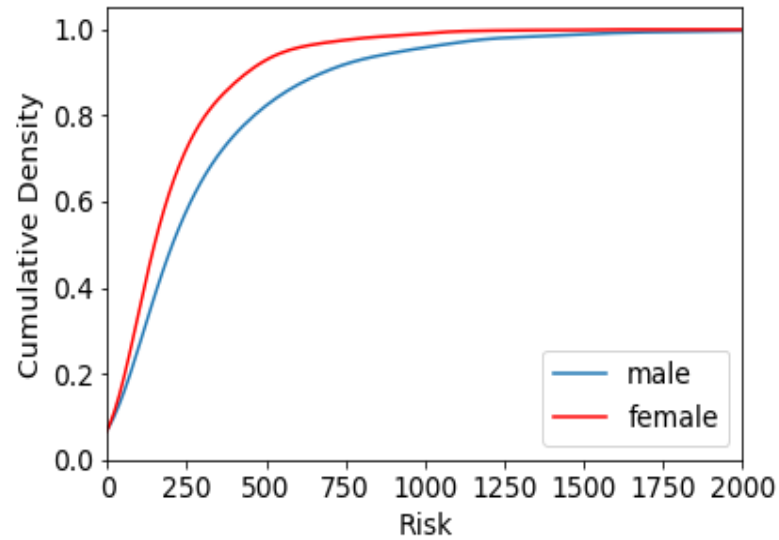
- **Price Optimization**

- Then we find the individualized profit-maximizing price of a single-product firm by solving a high-dimensional constrained optimization problem, utilizing the recent progress of optimization techniques Cotter et al. (2019).

- **Market Specification**

- Scenario 1: The outside option is uninsured – voluntary insurance in a monopoly
- Scenario 2: The outside option is another insurance product by other firms – compulsory insurance in a competitive market.
 - There exists another insurance product available at the price: $p_i^{out} = \delta \hat{c}_i^{out}$. $\delta > 1$ determines how competitive the market is.

Discrepancy between Male and Female



- Female consumers are:
 - safer drivers
 - charged less price
 - charged more markup

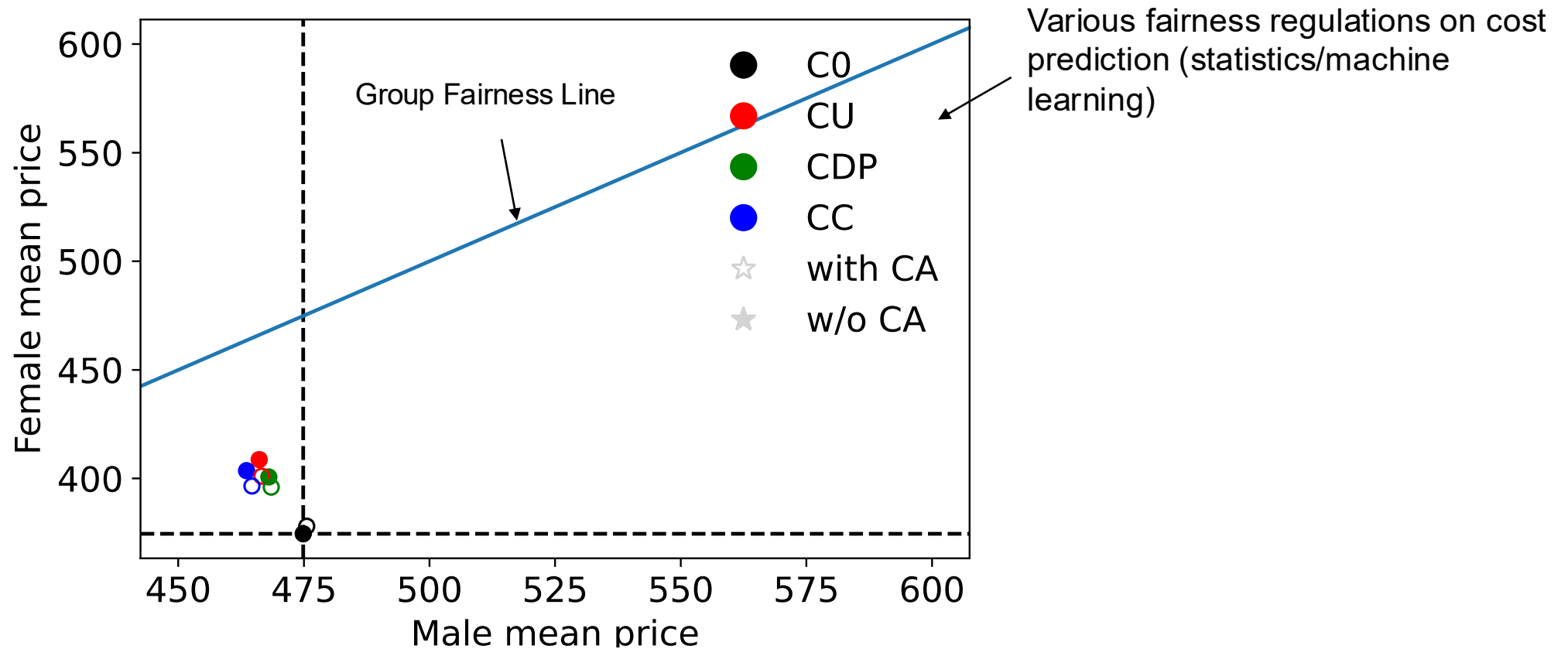
Which group is the "disadvantaged"?

First Finding

Is fair insurance pricing a fair machine learning problem?

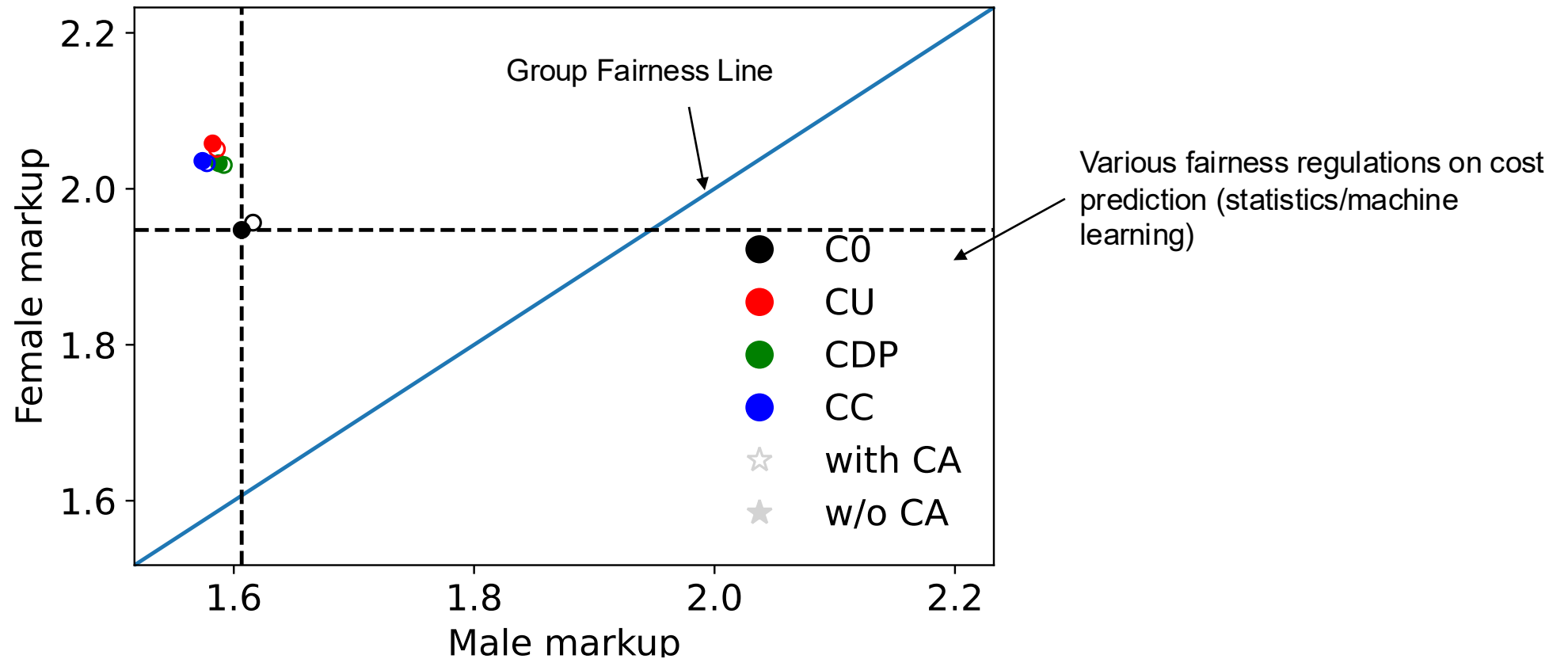
Fairness in machine learning (cost prediction) \neq Fairness in outcome (pricing)

Machine Learning Fairness is different from Outcome Fairness – **Market Price**



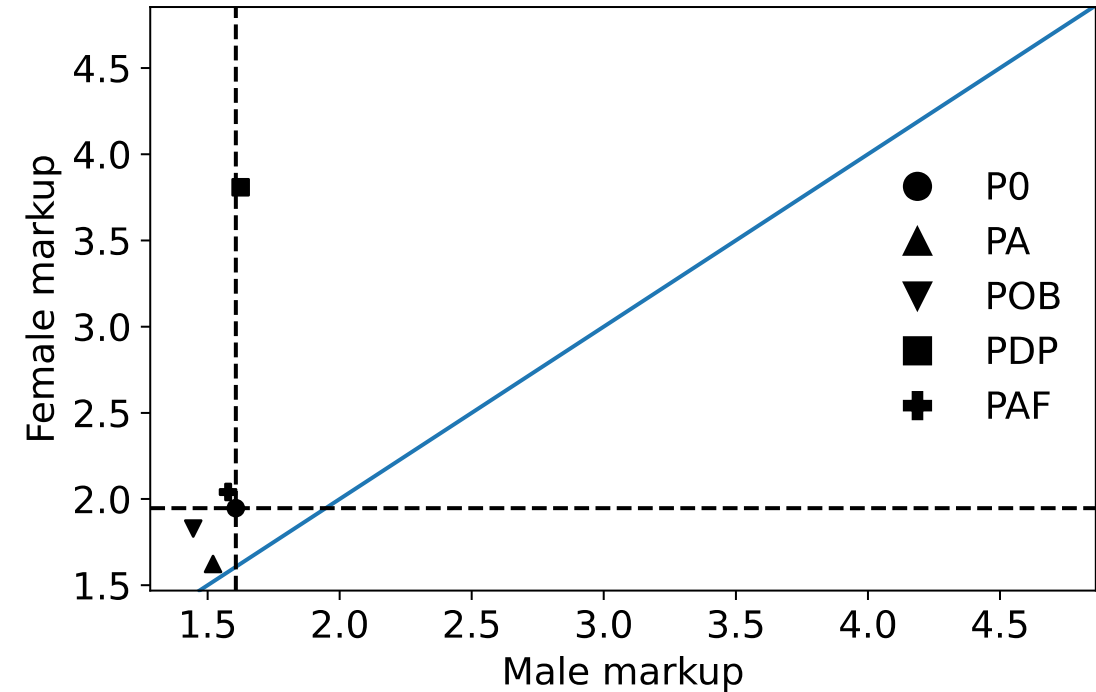
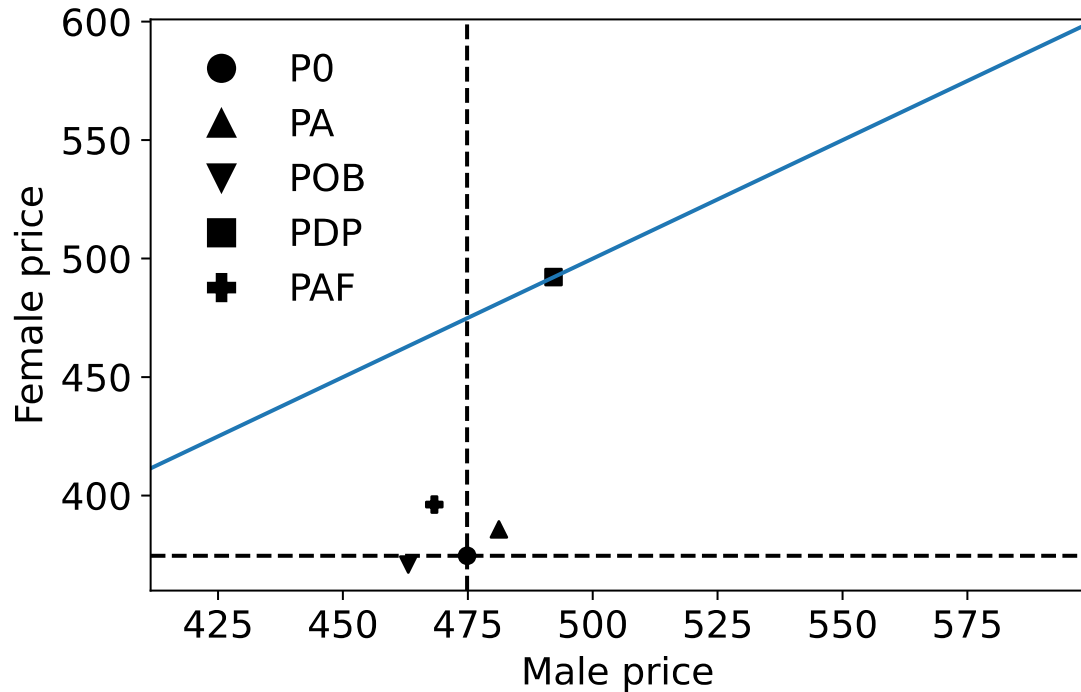
- Females have **lower** prices on average.
- Fair regulations on cost modelling (machine learning) **reduces the price gaps** among groups.

Machine Learning Fairness is different from Outcome Fairness -- Markup



- Females have **higher** markups on average.
- Fair regulations on cost modelling **increases the markup gaps** between gender groups.

Results - Regulations on Pricing



- PDP equalizes the prices among two gender groups and creates a huge markup gap.
- Which measure should we focus on: price or markup or else (e.g. loss ratio)?

Second Finding

How to evaluate the fairness criteria and choose a trade-off?

Welfare
Fairness-~~Accuracy~~ Trade-off

Our empirical results show that

- Decrease machine learning accuracy by only 0.5% can decrease 5% of profit and consumer welfare.

Small prediction accuracy drop can lead to big profit and welfare loss.

More Questions and Next Steps

1. Fairness beyond insurance pricing
2. How to overcome the difficulty that protected attributes are not collected?
3. How to deal with climate disasters and insurance affordability?
4. Can we trust interpretation tools in machine learning to explain fair decisions?
5. What level of transparency is needed to ensure fairness? Data, algorithms, human subjective decisions.
6. Can we have specific guidance to support informed decision-making regarding fairness criteria tailored for application contexts?
7. Call for industry-academia collaboration and better datasets!

