





# Fairness and discrimination in actuarial pricing

Arthur Charpentier<sup>1</sup> & Laurence Barry<sup>2</sup>

<sup>1</sup> Université du Québec à Montréal (UQAM) <sup>2</sup> Chaire PARI

Optimind, October 2022

# Agenda

- ▶ Barry and Charpentier (2022) The Fairness of Machine Learning in Insurance: New Rags for an Old Man?, *ArXiv:2205.08112* 
- ▶ Charpentier (2022) Insurance: Discrimination, Biases and Fairness, *Institut Louis Bachelier* 
- ▶ Grari et al. (2022) A fair pricing model via adversarial learning, *ArXiv:2202.12008*  



The collage contains the following elements:

- OPINIONS & DÉBATS** logo with a speech bubble.
- Book cover for **Insurance: Discrimination, Biases & Fairness** by Arthur Charpentier.
- Contents** table of contents for the book, listing chapters and page numbers.
- Summary** section.
- Text from the book, including a paragraph starting with "Insurance is a natural and the economic basis..." and another starting with "The insurance industry is a natural and the economic basis...".
- A small box at the bottom right with the text: "This book is published in the context of the Institut Louis Bachelier, Paris and Institut des Sciences de la Gestion, Université de Lausanne, Switzerland.".

# Notations

|                               |   |
|-------------------------------|---|
| $y \in \{0, 1\}$              | variable of interest (classically binary)             |
| $p \in \{0, 1\}$              | protected variable (sensitive)                        |
| $\mathbf{x} \in \mathbb{R}^d$ | 'explanatory' variables                               |
| $s \in [0, 1]$                | score, classically $s = s(\mathbf{x}, p)$             |
| $\hat{y} \in \{0, 1\}$        | classifier, classically $\hat{y} = \mathbf{1}(s > t)$ |

## Fairness Through Unawareness, Kusner et al. (2017)

Protected attribute  $p$  is not explicitly used in decision function  $\hat{y}$ .

## Ethics, Fairness and Discrimination

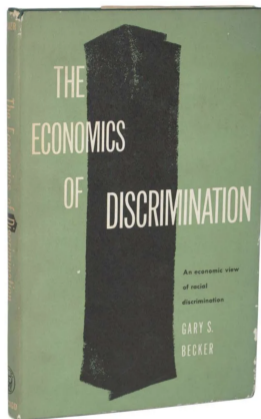
- ▶ **Accuracy** :  $\pi(\mathbf{x}) = \mathbb{E}_{\mathbb{P}}[Y|\mathbf{X} = \mathbf{x}]$  ( $\mathbb{P}$  historical probability) (*is*)
- ▶ **Fairness** :  $\pi^*(\mathbf{x}) = \mathbb{E}_{\mathbb{P}^*}[Y|\mathbf{X} = \mathbf{x}]$  ( $\mathbb{P}^*$  targeted probability) (*ought*, Hume (1739))
- ▶ “*Technology is neither good nor bad; nor is it neutral*” , Kranzberg (1986)
- ▶ “*Machine learning won't give you anything like gender neutrality 'for free' that you didn't explicitly ask for*”, Kearns and Roth (2019)
- ▶ “*at the core of insurance business lies discrimination between risky and non-risky insureds*”, Avraham (2017)
- ▶ Discrimination could be described as “*actuarially fair*”, therefore characterizing “*fair discrimination*” when the difference in premium reflects a difference in risk, Meyer and Rothstein (2004),
- ▶ “*Insurance rates are color-blind and solely based on risk*”,  
(Property Casualty Insurers Association of America, 2015 CFA study)  
cited in <https://evolutionofraceandinsurance.org/2000s>

## “actuarial fair discrimination” ?

Bohren et al. (2019) on statistical discrimination, or efficient discrimination, as in Becker (1957) (inspired by Edgeworth (1922) up to Phelps (1972))

Becker (2005) says “*if young Moslem Middle Eastern males were in fact much more likely to commit terrorism against U.S. than were other groups, putting them through tighter security clearance would reduce current airport terrorism*” ,

“*racial profiling*” is “*effective*”, even though “*such profiling is ‘unfair’ to the many young male Moslems who are not terrorists, and to the many minority shoppers who are honest ...*”



## Protected Attributes ?

|                    | CA | HI | GA | NC | NY | MA | PA | FL | TX | AL | ON | NB | NL | QC |
|--------------------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| Gender             | ×  | ×  | ●  | ×  | ●  | ×  | ×  | ●  | ●  | ●  | ●  | ×  | ×  | ●  |
| Age                | ×  | ×  | ●  | ×  | ●  | ×  | ●  | ●  | ●  | ●  | ●  | ×  | ×  | ●  |
| Driving experience | ●  | ×  | ●  | ●  | ●  | ●  | ●  | ●  | ●  | ●  | ●  | ●  | ●  | ●  |
| Credit history     | ×  | ×  | ●  | ●  | ●  | ×  | ●  | ●  | ●  | ×  | ×  | ●  | ×  | ●  |
| Education          | ×  | ×  | ×  | ×  | ×  | ×  | ●  | ●  | ●  | ●  | ●  | ●  | ●  | ●  |
| Occupation         | ×  | ×  | ×  | ●  | ×  | ×  | ●  | ●  | ●  | ●  | ●  | ●  | ●  | ●  |
| Employment status  | ×  | ×  | ×  | ●  | ×  | ×  | ●  | ●  | ●  | ●  | ●  | ●  | ●  | ●  |
| Marital status     | ●  | ×  | ●  | ●  | ●  | ×  | ●  | ●  | ●  | ●  | ●  | ●  | ●  | ●  |
| Housing situation  | ×  | ×  | ●  | ●  | ●  | ×  | ●  | ●  | ●  | ×  | ×  | ●  | ●  | ●  |
| Address/ZIP code   | ●  | ●  | ●  | ●  | ●  | ●  | ●  | ●  | ●  | ×  | ×  | ●  | ●  | ●  |
| Insurance history  | ●  | ●  | ●  | ●  | ●  | ●  | ●  | ●  | ●  | ●  | ●  | ●  | ●  | ●  |

CA: Californie, HI: Hawaii, GA: Georgia, NC: Caroline du nord, NY: New York, MA: Massachusetts, PA: Pennsylvanie, FL: Floride, TX: Texas, AL: Alberta, ON: Ontario, NB: Nouveau-Brunswick, NL: Terre-Neuve-et-Labrador, QC: Québec

|                               |   |
|-------------------------------|---|
| $y \in \{0, 1\}$              | variable of interest (classically binary)             |
| $p \in \{0, 1\}$              | protected variable (sensitive)                        |
| $\mathbf{x} \in \mathbb{R}^d$ | 'explanatory' variables                               |
| $s \in [0, 1]$                | score, classically $s = s(\mathbf{x}, p)$             |
| $\hat{y} \in \{0, 1\}$        | classifier, classically $\hat{y} = \mathbf{1}(s > t)$ |

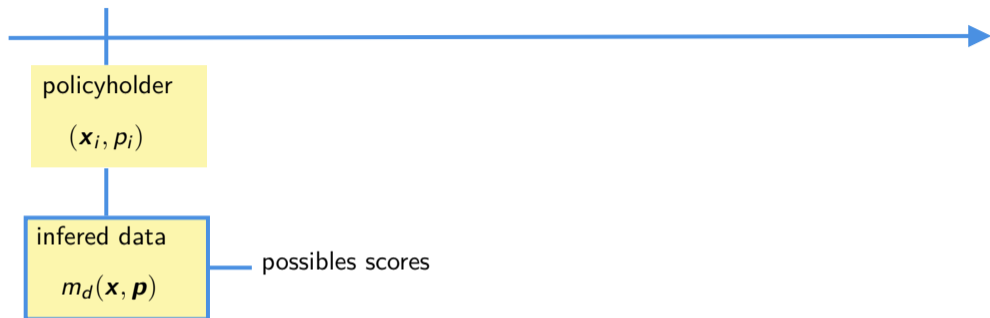
## Fairness Through Unawareness, Kusner et al. (2017)

Protected attribute  $p$  is not explicitly used in decision function  $\hat{y}$ .

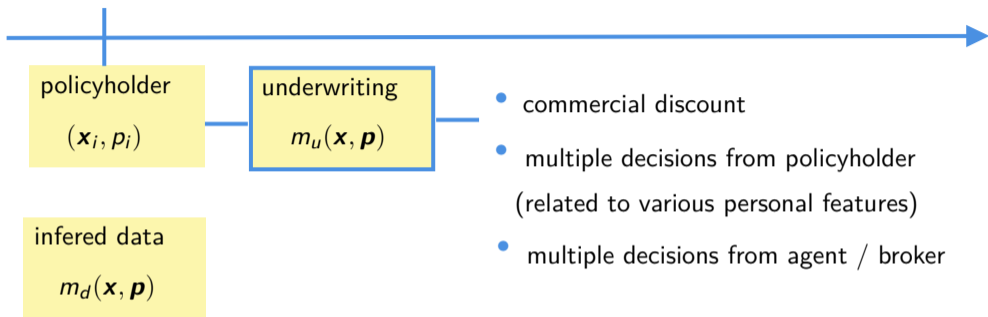
## Data & Models II



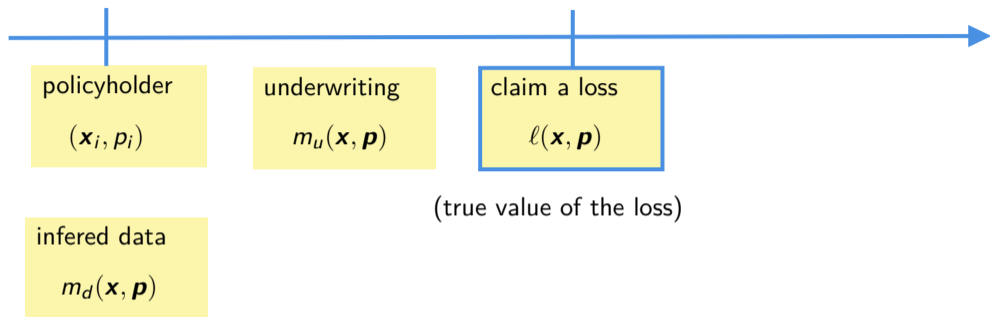
## Data & Models III



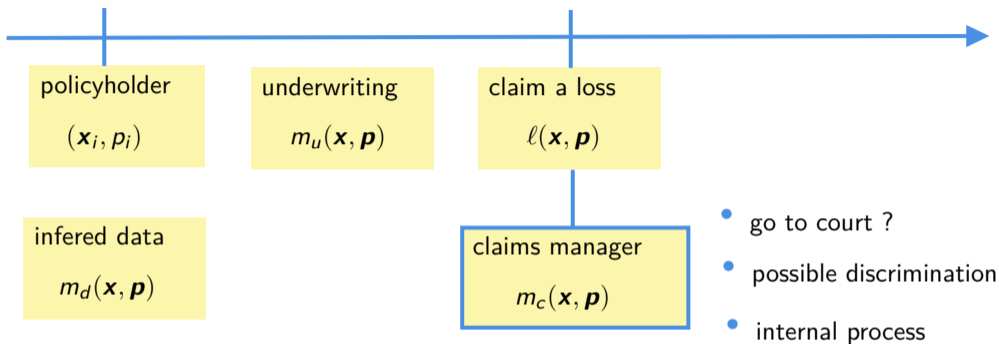
## Data & Models IV



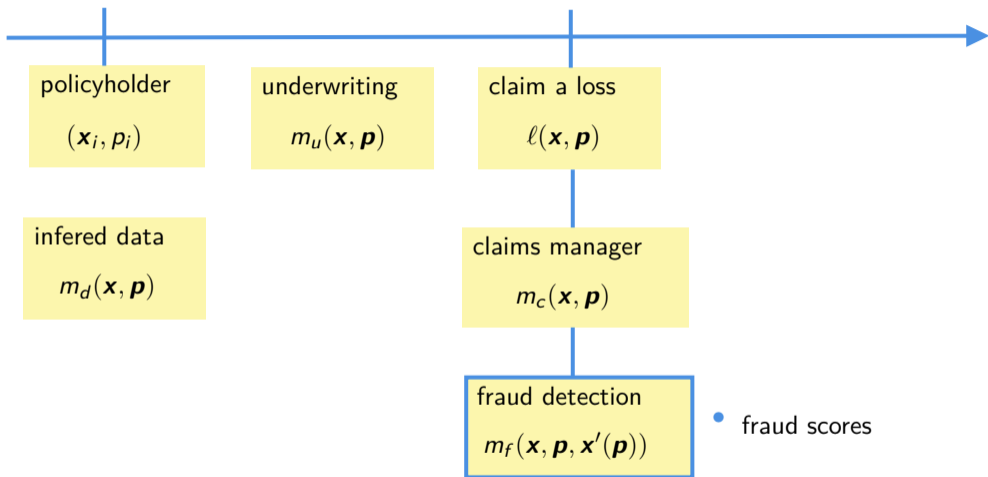
# Data & Models V



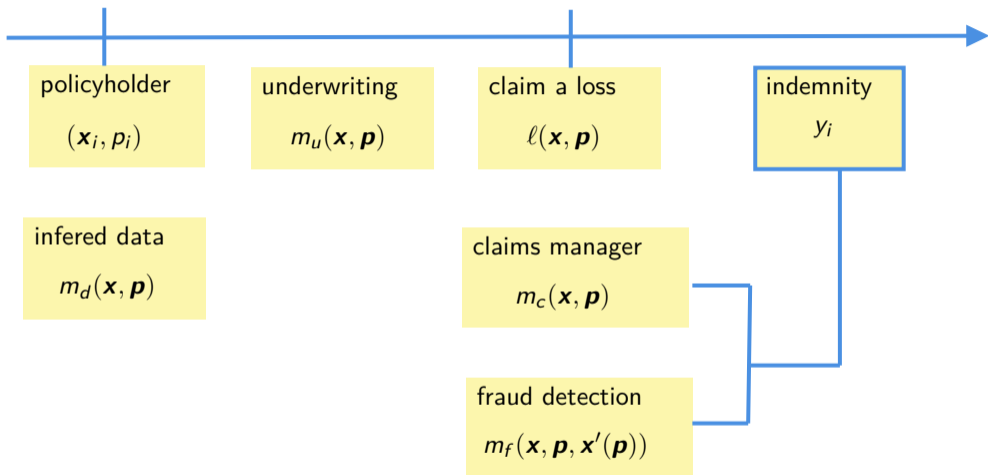
# Data & Models VI



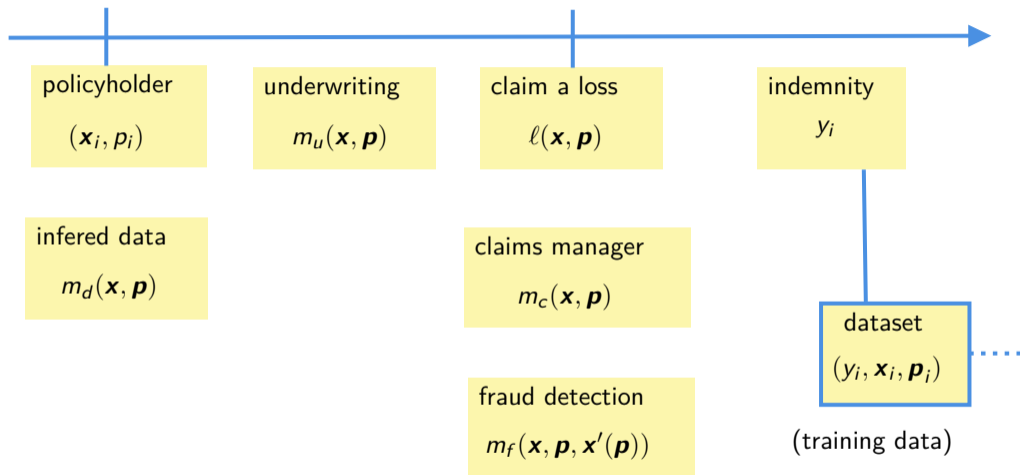
# Data & Models VII



# Data & Models VIII



# Data & Models IX



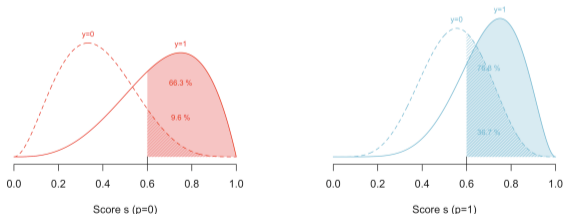
# Defining Group Fairness I

**Demographic Parity**, (Corbett-Davies et al. (2017), Agarwal (2021))

Decision function  $\hat{y}$  satisfies demographic parity if  $\hat{Y} \perp\!\!\!\perp P$ , i.e.

$$\mathbb{P}[\hat{Y} = y | P = 0] = \mathbb{P}[\hat{Y} = y | P = 1], \forall y \text{ or } \mathbb{E}[\hat{Y} | P = 0] = \mathbb{E}[\hat{Y} | P = 1]$$

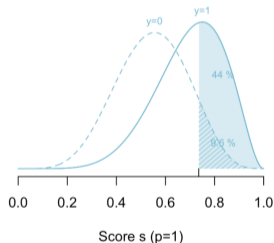
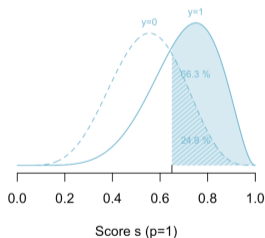
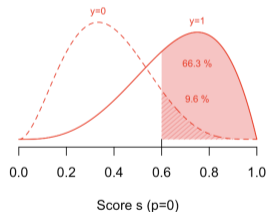
We can compare  $s(\mathbf{X})$  conditional on  $Y$ , but also on  $P$



# Defining Group Fairness II

## Equal Opportunity, Hardt et al. (2016)

True positive parity  $\mathbb{P}[\hat{Y} = 1|P = 0, Y = 1] = \mathbb{P}[\hat{Y} = 1|P = 1, Y = 1]$  or false positive parity  $\mathbb{P}[\hat{Y} = 1|P = 0, Y = 0] = \mathbb{P}[\hat{Y} = 1|P = 1, Y = 0]$



# Defining Group Fairness III

|                                       |                              |   |                                  |
|---------------------------------------|------------------------------|---|----------------------------------|
| <i>statistical parity</i>             | Dwork et al. (2012)          | $\mathbb{P}[\hat{Y} = 1   P = p] = \text{cst}, \forall p$             | independence                     |
| <i>conditional statistical parity</i> | Corbett-Davies et al. (2017) | $\mathbb{P}[\hat{Y} = 1   P = p, X = x] = \text{cst}_x, \forall p, x$ | $\hat{Y} \perp\!\!\!\perp P$     |
| <i>equalized odds</i>                 | Hardt et al. (2016)          | $\mathbb{P}[\hat{Y} = 1   P = p, Y = y] = \text{cst}_y, \forall p, y$ | separation                       |
| <i>equalized opportunity</i>          | Hardt et al. (2016)          | $\mathbb{P}[\hat{Y} = 1   P = p, Y = 1] = \text{cst}, \forall p$      |                                  |
| <i>predictive equality</i>            | Corbett-Davies et al. (2017) | $\mathbb{P}[\hat{Y} = 1   P = p, Y = 0] = \text{cst}, \forall p$      | $\hat{Y} \perp\!\!\!\perp P   Y$ |
| <i>balance (positive)</i>             | Kleinberg et al. (2017)      | $\mathbb{E}[S   P = p, Y = 1] = \text{cst}, \forall p$                | $S \perp\!\!\!\perp P   Y$       |
| <i>balance (negative)</i>             | Kleinberg et al. (2017)      | $\mathbb{E}[S   P = p, Y = 0] = \text{cst}, \forall p$                |                                  |
| <i>conditional accuracy equality</i>  | Berk et al. (2017)           | $\mathbb{P}[Y = y   P = p, \hat{Y} = y] = \text{cst}_y, \forall p, y$ | sufficiency                      |
| <i>predictive parity</i>              | Chouldechova (2017)          | $\mathbb{P}[Y = 1   P = p, \hat{Y} = 1] = \text{cst}, \forall p$      |                                  |
| <i>calibration</i>                    | Chouldechova (2017)          | $\mathbb{P}[Y = 1   P = p, S = s] = \text{cst}_s, \forall p, s$       | $Y \perp\!\!\!\perp P   \hat{Y}$ |
| <i>well-calibration</i>               | Chouldechova (2017)          | $\mathbb{P}[Y = 1   P = p, S = s] = s, \forall p, s$                  |                                  |
| <i>accuracy equality</i>              | Berk et al. (2017)           | $\mathbb{P}[\hat{Y} = Y   P = p] = \text{cst}, \forall p$             |                                  |
| <i>treatment equality</i>             | Berk et al. (2017)           | $\frac{\text{FN}_p}{\text{FP}_p} = \text{cst}_p, \forall p$           |                                  |

# Individual Fairness

Consider some distances  $D$  on  $\{0, 1\} \times \{0, 1\}$  or  $[0, 1] \times [0, 1]$ , and  $d$  on  $\mathbb{R}^p \times \mathbb{R}^p$ ,

**Lipschitz property**, Duivesteijn and Feelders (2008)

$$D(\hat{y}_i, \hat{y}_j) \text{ or } D(s_i, s_j) \leq d(\mathbf{x}_i, \mathbf{x}_j), \quad \forall i, j = 1, \dots, n.$$

**Counterfactual fairness**, Kusner et al. (2017) If the prediction in the real world is the same as the prediction in the counterfactual world where the individual would have belonged to a different demographic group, we have counterfactual equity, i.e.

$$\mathbb{P}[Y_{P \leftarrow p}^* = y | \mathbf{X} = \mathbf{x}] = \mathbb{P}[Y_{P \leftarrow p'}^* = y | \mathbf{X} = \mathbf{x}], \quad \forall p', \mathbf{x}, y.$$

## Penalizing to get a more fair pricing model

Inspired by [Goodfellow et al. \(2018\)](#), to avoid un-fairness: penalize to a dependence measure between  $\hat{y}$  and  $p$ , e.g. for **demographic parity**

$$\operatorname{argmin}_{\theta} \{ \mathcal{L}(h_{\theta}(\mathbf{x}), y) + \lambda \operatorname{corr}(\hat{y}, p) \}$$

or for **equalized odds**, if  $y \in \{0, 1\}$

$$\operatorname{argmin}_{\theta} \{ \mathcal{L}(h_{\theta}(\mathbf{x}), y) + \lambda_0 \operatorname{corr}(\hat{y}, p | y = 0) + \lambda_1 \operatorname{corr}(\hat{y}, p | y = 1) \}$$

(that could be extended to more general  $y$ 's, see [Grari et al. \(2022\)](#), possibly continuous, with more interesting dependence measures such as the maximal correlation)

## References I

- Agarwal, S. (2021). Trade-offs between fairness and interpretability in machine learning. In *IJCAI 2021 Workshop on AI for Social Good*.
- Avraham, R. (2017). Discrimination and insurance. In Lippert-Rasmussen, K., editor, *Handbook of the Ethics of Discrimination*, pages 335–347. Routledge.
- Barry, L. and Charpentier, A. (2020). Personalization as a promise: Can big data change the practice of insurance? *Big Data & Society*, 7(1):2053951720935143.
- Becker, G. S. (1957). *The economics of discrimination*. University of Chicago press.
- Becker, G. S. (2005). Is ethnic and other profiling discrimination? *The Becker-Posner Blog*, 01-23-2005.
- Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. (2017). A convex framework for fair regression. *arXiv*, 1706.02409.
- Bohren, J. A., Haggag, K., Imas, A., and Pope, D. G. (2019). Inaccurate statistical discrimination: An identification problem. Technical report, National Bureau of Economic Research.
- Charpentier, A. (2022). *Insurance: biases, discrimination and fairness*. Institut Louis Bachelier.
- Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.

## References II

- Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. *arXiv*, 1701.08230.
- Duivesteijn, W. and Feelders, A. (2008). Nearest neighbour classification with monotonicity constraints. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 301–316. Springer.
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.
- Edgeworth, F. Y. (1922). Equal pay to men and women for equal work. *The Economic Journal*, 32(128):431–457.
- Goodfellow, I., McDaniel, P., and Papernot, N. (2018). Making machine learning robust against adversarial inputs. *Communications of the ACM*, 61(7):56–66.
- Gari, V., Charpentier, A., Lamprier, S., and Detyniecki, M. (2022). A fair pricing model via adversarial learning. *ArXiv*, 2202.12008.
- Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323.
- Hume, D. (1739). *A Treatise of Human Nature*. Cambridge University Press Archive.

## References III

- Kearns, M. and Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.
- Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2017). Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, 133(1):237–293.
- Kranzberg, M. (1986). Technology and history: "kranzberg's laws". *Technology and culture*, 27(3):544–560.
- Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4066–4076. NIPS.
- Meyer, R. B. and Rothstein, M. A. (2004). The insurer perspective. In Rothstein, M. A., editor, *Genetics and Life Insurance*, pages 27–47. MIT Cambridge Massachusetts,.
- Phelps, E. S. (1972). The statistical theory of racism and sexism. *The american economic review*, 62(4):659–661.