le cnam

On some issues related to the fairness of algorithms

gilbert.saporta@cnam.fr

Outline

1. The triumph of the black boxes in the 2000s
2. Societal implications
3. Interpretability and explainability
4. New requirements: interpretable and causal models
5. Fairness: a statistical problem?
6. Conclusion and perspectives

1. The triumph of black boxes in the 2000s

- Model factories
  - Systems that automatically generate predictive models with little or no human intervention. e.g.: simultaneous sales forecasts of thousands of items
  - 2010 KMF : Kxen Modeling Factory for automated risk score generation

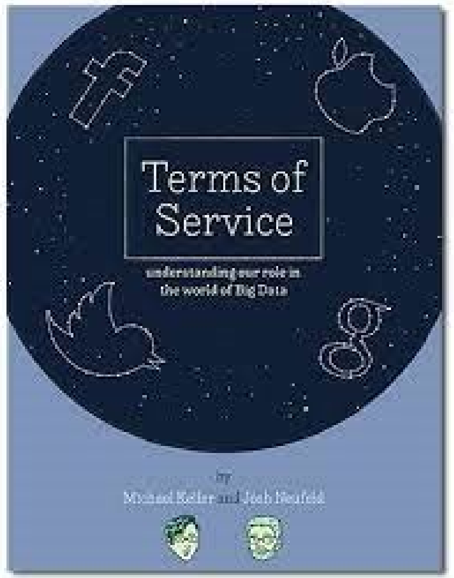
A French-American startup founded in 1998, acquired by SAP in 2013, based on an original idea by Léon Bottou using the Vapnik-Cervonenkis theory on structural risk minimization.

Is the "why" so important?

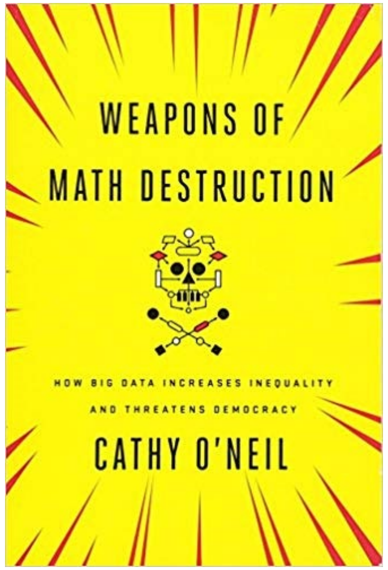
- In everyday life, we trust many processes that we do not understand: cars, television, smartphones, weather forecasts. No matter if black boxes are used.
- But when certain decisions have implications on our lives: health, employment, money, etc., the right to an explanation is necessary.

2. Societal implications

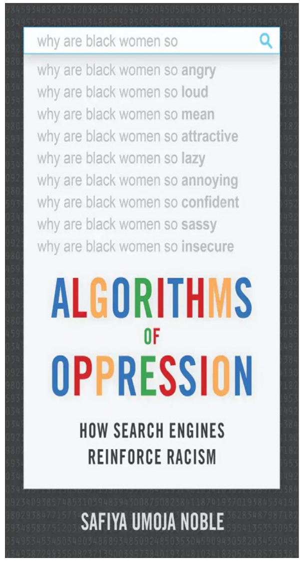
- A denunciation literature motivated by unethical applications of Machine Learning in massive automatic decisions on individuals: e.g. loan allocation, predictive justice, recruitment....



2014



2016



2018

THE WALL STREET JOURNAL.

Subs

S

English Edition ▾ | Print Edition | Video | Podcasts | Latest Headlines

Home World U.S. Politics Economy Business Tech Markets Opinion Life & Arts Real Estate WSJ. Magazi

DIGITS

Google Mistakenly Tags Black People as 'Gorillas,' Showing Limits of Algorithms

By Alistair Barr

Updated July 1, 2015 3:41 pm ET

Black programmer Jacky Alciné said on Twitter that the new Google Photos app had tagged photos of him and a friend as gorillas.

ILLUSTRATION: JACKY ALCINÉ AND TWITTER

- *The increasing use of algorithms to make eligibility decisions must be carefully monitored for potential discriminatory outcomes for disadvantaged groups, even absent discriminatory intent.*

  Executive Office of US President: Big Data: Seizing Opportunities and Preserving Values, https://www.hsdl.org/?view&did=752636 (2014)

- *Through 2022, 85 percent of AI projects will deliver erroneous outcomes due to bias in data, algorithms or the teams responsible for managing them.*

https://www.gartner.com/en/newsroom/press-releases/2018-02-13-gartner-says-nearly-half-of-cios-are-planning-to-deploy-artificial-intelligence

The right to an explanation

- **EU High Level Expert Group on AI (2019)**

- Respect for human autonomy
- Prevention of any harm
- Fairness
- Explainability

- **OECD Council Recommendation on AI (2019)**

- Inclusive growth, sustainable development and well-being
- Human-centred values and fairness
- Transparency and explainability
- Robustness, security and safety

Codes, regulations and ethics statements

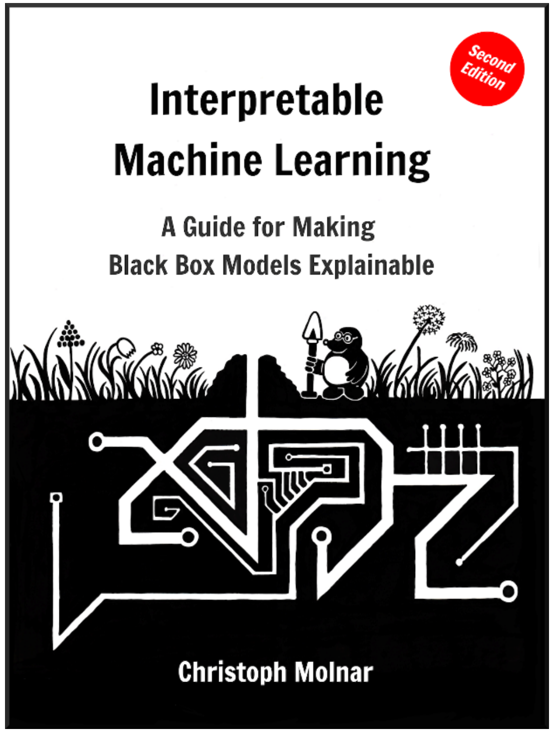
- EU General Data Protection Regulation (2016) https://gdprinfo.eu
- Montréal Declaration for a Responsible Development of Artificial Intelligence (2017) https://www.montrealdeclaration-responsibleai.com/
- NYC Local Law on Automatic Decision Systems (2018) https://legistar.council.nyc.gov/View.ashx?M=F&ID=5828157&GUID=4A07389A-0FE9-432B-8130-D9E1821C82C4
- OECD Recommendation of the Council on Artificial Intelligence (2019), https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449
- EU Ethics guidelines for trustworthy AI (2019) https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
- Council of Europe. Towards regulation of AI systems (2020) https://www.coe.int/en/web/artificial-intelligence/cai
- UNESCO Recommendation on the ethics of artificial intelligence (2021) https://en.unesco.org/artificial-intelligence/ethics
- EU Digital Services Act, Digital Market Act (2022) https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package

Expected properties of the algorithms

- **Transparency:** on the purpose, structure and underlying actions of the algorithms used
- **Responsability:** companies should be held accountable for the results of their programmed algorithms.
- **Auditability:** Describes the ability to evaluate algorithms, models, and datasets; to analyze the operation, results, and effects, even unexpected, of AI systems.
- **Fairness:** if its results are independent of variables considered sensitive, such as characteristics of individuals that should not be correlated with the outcome (gender, ethnicity, sexual orientation, disability, etc.)

3. Interpretability and explainability of models

- Terms often used interchangeably
- Interpretability
  - Refers to **simple** and transparent algorithms: logical models (trees, ...), linear (sparse, ...), knn.
- Explainability
  - the ability to explain or present in terms understandable to a human being
  - Generally **post-hoc** (open the black box)
    - Local or global
    - Specific or **agnostic**



2022

3.1 Measures of variables importance

3.1.1 Specific methods

- It is often believed that simple models, such as linear or logistic regression, are easily interpreted.
- Generally untrue!
- Except in the case of orthogonal designs, the parameter values hardly reflect the importance of the variables.

- More than 14 methods to quantify the importance of variables in linear models!(Grömping, 2015, Wallard, 2015)

R package `relaimpo`

b's (not normalized)
Joint contribution (not normalized)
Squared semipartial correlations
Squared raw correlations
Squared standardized b's
Sequential SS, from left to right
Sequential SS, from right to left
Pratt
CAR scores/Gibson
Green et al.
Fabbris
Genizi/Johnson
LMG
PMVD

3.1.2 Agnostic methods

permutation variable importance (Breiman, 2001)

- Introduced for random forests as the increase of the prediction error of the model when shuffling the predictor values.
- Easy to understand approach, taking into account the interactions
- Must be repeated and averaged
- Importances are not additive
- Can lead to physically impossible unit pairs and outliers.

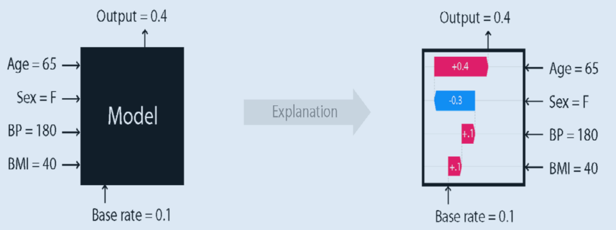
Shapley value

Inspired by game theory(Lundberg & Lee, 2017)

- Nice mathematical properties
  - Including additivity and uniqueness under certain conditions.
  - Allows to decompose an individual prediction (local values)
  - Global importance of a predictor: average of the local values on all n units

Python libraries

SHAP



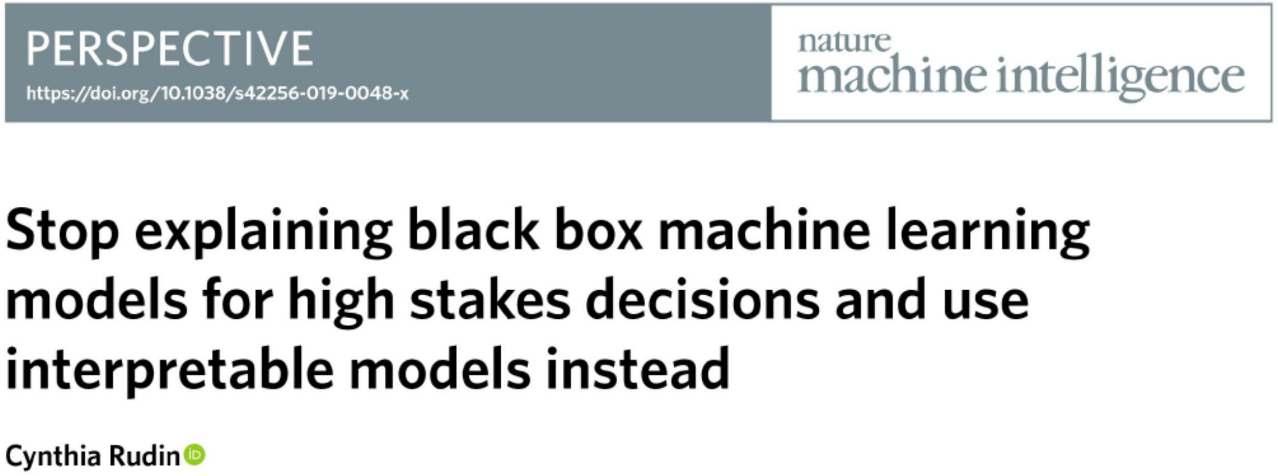
shapash

3.2 Surrogate models

"A surrogate model is an interpretable model that is trained to approximate the predictions of a black box model" (Molnar, 2020)

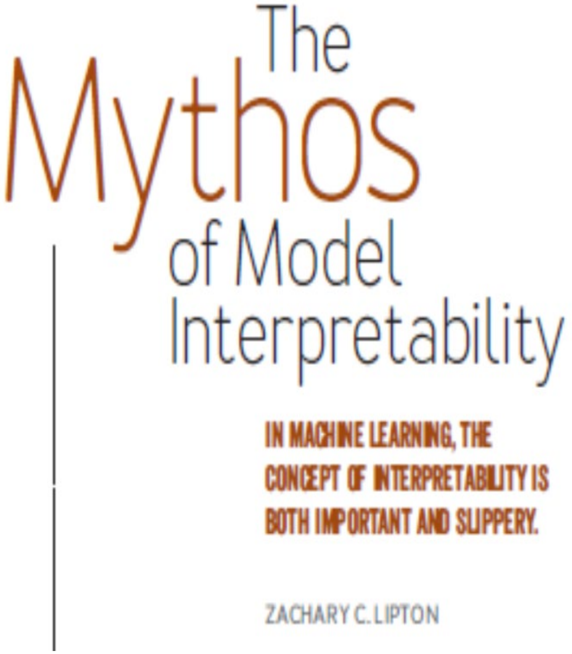
- Can be global or local, agnostic or specific.
- Agnostic means that it can be applied to any learning model.
- A surrogate model tries to approximate the black box model, not to fit the data.
- Trees, linear models are the preferred alternative models.
- A popular approach : **LIME** (**L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations)

4. New requirements: interpretable and causal models


PERSPECTIVE
https://doi.org/10.1038/s42256-019-0048-x
nature machine intelligence

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin



Interpretability and causality

- Measuring the importance of a variable does not answer this question: what would be the answer if one or more predictors were changed intentionally or unintentionally?
  - It is often absurd to measure the effect of a variable "all other things being equal "..
  - Changing x_j may change the values of other predictors if they are causally related.

- Regression, ML models are not causal, but are often used as if they were, resulting in many disappointments..
- Seeing is not doing (Pearl & Mackenzie, 2018)

$$P(Y \mid X = x) \neq P(Y \mid do(X = x))$$

- Confusion between correlation and causation
- Difficult to infer causality from observational data.
  - Propensity score matching (Rosenbaum et Rubin, 1983)
  - Bayesian network learning
- Need for experiments

5. Fairness: a statistical problem?

- Motivated by discriminatory treatment of groups of people: gender, race etc. .
  - Binary classification
    - Decisions on remand and risk of recidivism (Wang et al. 2020)
    - Hiring
- An avalanche of articles and conferences over the past 4 years
  - FairWare 2018 https://fairware.cs.umass.edu/
  - ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT) since 2018 https://facctconference.org/

FAccT 2021
Toronto, Canada

- Set of descriptors V=(A,X) where A are sensitive variables (protected groups); Y binary outcome to predict, D decision or predicted outcome D=f(V)
  - In the U.S., sensitive categories are legally protected groups. Federal law makes it illegal to discriminate on the basis of: race, color, national origin, religion, sex, disability, age (40+), citizenship status, genetic information.
  - In California: 18 protected groups https://www.senate.ca.gov/content/protected-classes

5.1 Measuring fairness

		Definition
Statistical	3.1.1	Group fairness or statistical parity
	3.1.2	Conditional statistical parity
	3.2.1	Predictive parity
	3.2.2	False positive error rate balance
	3.2.3	False negative error rate balance
	3.2.4	Equalised odds
	3.2.5	Conditional use accuracy equality
	3.2.6	Overall accuracy equality
	3.2.7	Treatment equality
	3.3.1	Test-fairness or calibration
	3.3.2	Well calibration
	3.3.3	Balance for positive class
	3.3.4	Balance for negative class
Similarity-Based	4.1	Causal discrimination
	4.2	Fairness through unawareness
	4.3	Fairness through awareness
Causal Reasoning	5.1	Counterfactual fairness
	5.2	No unresolved discrimination
	5.3	No proxy discrimination
	5.4	Fair inference

Verma, Rubin 2018

More than 20 measures of algorithmic fairness!

New tools (open source)

- What If Tool (Google)

Playing with AI Fairness

Google's new machine learning diagnostic tool lets users try on five different types of fairness.

- AI Fairness 360 (IBM)

more than 70 metrics

- deon toolkit

"Adds an ethics checklist to your data science projects"

- Aequitas

Center for Data Science and Public Policy at University of Chicago

Measuring fairness without considering the outcome Y

- **Equality of decision measures**
  - Statistical parity or group equity if D is independent of A : $D \perp A$
  - Conditional demographic parity : $D \perp A$ knowing V
- **Metric fairness or individual fairness:** two individuals who are close according to V should be treated similarly.
  - Metric fairness implies **unawareness** if the metric considers only non-sensitive variables

- **Fairness through unawareness** :

D⊥A given X (non protected attributes) . The model should not explicitly use protected attributes A. Unawareness implies that people with the same x will be treated similarly.

  - A naïve approach might require that the algorithm should ignore all protected attributes such as race, color, religion, gender, disability, or family status. However, this idea of “fairness through unawareness” is ineffective due to the existence of redundant encodings, ways of predicting protected attributes from other features (Hardt et al., 2016)

Measuring fairness by considering the outcome Y

- **Equal precision or demographic parity**

$$P(Y=1 \mid A=a) = P(Y=1 \mid A=a')$$

- For example, we might want a medical diagnostic tool to be equally accurate for people of any race or gender. (Mitchell et al. 2021)
- Loan assessment: people across groups have the same chance of getting the loan

Considering both decision D and outcome Y

Confusion Matrix

	$Y=1$	$Y=0$	$P(Y=1\|D)$	$P(Y=0\|D)$
$D=1$	True positive	False positive	$P(Y=1\|D=1)$: Positive predictive value	$P(Y=0\|D=0)$: False discovery rate
$D=0$	False negative	True negative	$P(Y=1\|D=0)$: False omission rate	$P(Y=0\|D=0)$: Negative predictive value
$P(D=1\|Y)$	$P(D=1\|Y=1)$: True positive rate	$P(D=1\|Y=0)$: False positive rate		
$P(D=0\|Y)$	$P(D=0\|Y=1)$: False negative rate	$P(D=0\|Y=0)$: True negative rate		$P(D=Y)$: Accuracy

Mitchell et al., 2021

Conditioning according to the outcome

1. Equality of false positive rates (and therefore true negative rates)

$$P(D=1 \mid Y=0, A=a) = P(D=1 \mid Y=0, A=a')$$

$$D \perp A \text{ given } Y=0$$

2. Equality of true positive rates (and therefore false negative rates) equality of chances $D \perp A \text{ given } Y=1$

These two pairs reflect a fairness notion that people with the same outcome should be treated the same, regardless of sensitive group membership. (Mitchell et al.)

- 3. Both: **separation or equalization of opportunities**

  Ex. Among all people who will not default, they have the same chance of getting the loan. Among people who will default, they have the same chance of being rejected.

Conditioning according to the decision (decision maker's point of view)

1. Equality of predicted value

$$P(Y=0 \mid D=1, A=a) = P(Y=0 \mid D=1, A=a')$$

$$Y \perp A \text{ given } D=1$$

2. Equality of positive predictive value (and therefore equality of false discovery rate) is **predictive parity** : Y ⊥ A given D=1

- *Ex. Among all people who are given a loan, across groups there is the same proportion of people who will not default (equal chance of success given acceptance).*

3. Both: **sufficiency,** which means that people who underwent the same decision would have had similar outcomes, regardless of the group.

5.2 Impossibility theorems

- The above definitions are generally not compatible mathematically or morally. For example, separation and sufficiency cannot occur simultaneously.
- Group equity and individual equity are generally incompatible:
  - *Applied to the case of college admissions, for example, group equity would require that admission rates be equal for protected attributes (gender, etc.), while individual equity would require that each person be assessed independently of gender.* (Bertail et al. 2019)

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) **and the controversy with ProPublica**

- ProPublica: COMPAS fails to meet **equal false positive rates** by race: Among defendants who were not rearrested, black defendants were twice as likely to be misclassified as high risk. They described the tool as biased against blacks
- According to COMPAS, they meet **equal positive predictive values**: among so-called high-risk individuals, the proportion of defendants who were rearrested is approximately the same, regardless of race.
- Both definitions of fairness can only be satisfied when (a) the recidivism rate and score distribution are the same for all racial groups or (b) some groups are not affected (e.g., whites are never rearrested).
- The definitions of equity defended by the different sides of the debate cannot be achieved simultaneously. (Mitchell et al., 2021)

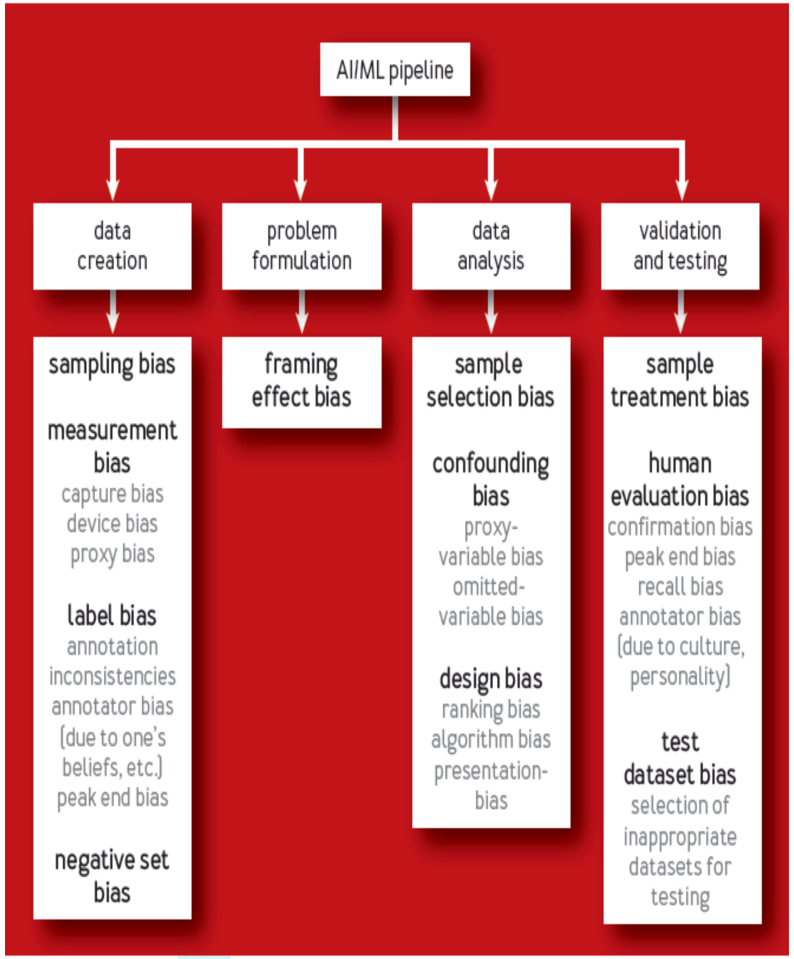
5.3 Operational issues

- The operational character of these measures is also problematic, not only because the **outcome** will often **not be observable** until well after the decision, but especially because of a **counterfactual** problem since in some cases the decision prohibits observation: we will never know whether a refused loan would have been repaid.
- In other cases, such as the selection of candidates for a job, the variable Y that would consist in determining whether the recruited candidate does the job well is not even observable. We don't know the ground truth and the algorithms only automate previous processes.

5.4 The "biases" of algorithms

- In most cases, algorithms simply reproduce the biases of learning data and human decisions.
- Statistical bias
  - Non-representative sample
  - Missing data and selection bias (e.g., loan applications)
  - Traditional remedies: reweighting

FIGURE 1: TAXONOMY OF BIAS TYPES ALONG THE AI PIPELINE

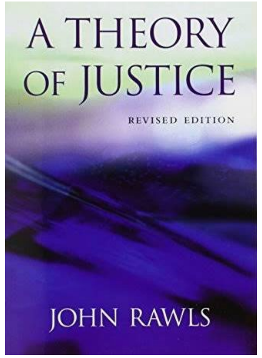


Srinivasan, R., & Chander, A. (2021).

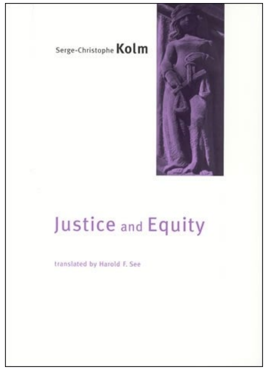
- Omitted variable bias
  - In the absence of a proxy, the omission of an important predictor in a model usually leads to erroneous results and is difficult to detect.
  - The omission of a variable can lead to an inversion of an association in the whole population compared to sub-populations
    - Simpson's paradox if the omitted variable is categorical, Berkson's or Lord's if the variable is continuous.
  - Measurement and technological biases: facial recognition fails to recognize people of color as accurately as it does white people.
- Historical biases, stereotypes and societal, cultural and cognitive prejudices, ...
  - Crime rates reflect unequal social structures and also inequalities in judgments
  - Data that are representative but reproduce inequalities (women's wages).

6. Conclusion and perspectives

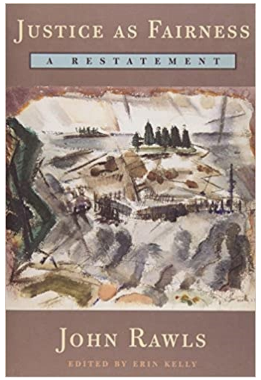
- Transparency and causality do not guarantee fairness
- No single measure of fairness
- Algorithm biases are often reproductions of previous biases
- Algorithmic equity must be placed in a more general framework of philosophy and political economy.



1971



1998



2001

References

- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Chatila, R. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82-115.
- Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*
- Dataiku (2022). *Black-Box vs. Explainable AI: How to Reduce Business Risk and Infuse Transparency.* https://content.dataiku.com/black-box-vs-explainable-ai/black-box-vs-explainable-ai
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214-226).
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5), 1-42.
- Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI Magazine*, *40*(2), 44-58.
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems*, 29, 3315-3323

- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5), 1-42.
- Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI Magazine*, *40*(2), 44-58.
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems*, 29, 3315-3323
- Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2017). Interpretable & explorable approximations of black box models. *arXiv preprint arXiv:1707.01154*.
- Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31-57.
- Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, *8* (pp.141-163)
- Molnar, C. (2022). *Interpretable machine learning , A Guide for Making Black Box Models Explainable*, https://christophm.github.io/interpretable-ml-book.
- Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592*.
- Pearl, J., & Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic Books.

- Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. MIT press.
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215
- Rudin, C., & Radin, J. (2019). Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition. *Harvard Data Science Review*, *1*(2).
- Schölkopf, B. (2019). Causality for machine learning. *arXiv preprint arXiv:1911.10500*.
- Srinivasan, R., & Chander, A. (2021). Biases in AI Systems: A survey for practitioners. *Queue*, *19*(2), 45-64
- Tsamados, A., Aggarwal, N., Cowls, J. et al. (2021). The ethics of algorithms: key problems and solutions. *AI & Society*
- Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)* (pp. 1-7).

- Vapnik, V. (2006). *Estimation of Dependences Based on Empirical Data*, 2nd edition, Springer
- Varian, H. (2014). Big Data: New Tricks for Econometrics, *Journal of Economic Perspectives*, 28, 2, 3–28
- Varian, H. (2016). Causal inference in economics and marketing, *PNAS* , 113, 7310-7315
- Wallard, H. (2015). Using explained variance allocation to analyse importance of predictors. In *16th ASMDA conference proceedings* (Vol. 30), 1043-1054,
- Wang, C., Han, B., Patel, B., Mohideen, F., & Rudin, C. (2020). In Pursuit of Interpretable, Fair and Accurate Machine Learning for Criminal Recidivism Prediction. *arXiv preprint arXiv:2005.04176*