HUNGRY ECOCITIES

A S+T+ARTS RESIDENCIES PROJECT

S+T+ARTS

European Commission

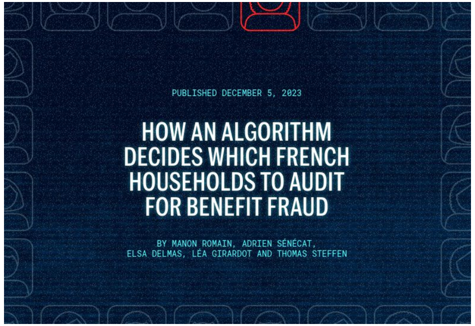
AI Bias, Fairness and Ethics: An Overview

Speaker: Kristen Scott

KU Leuven

This work has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie Actions grant agreement number 860630).

Algorithmic harms

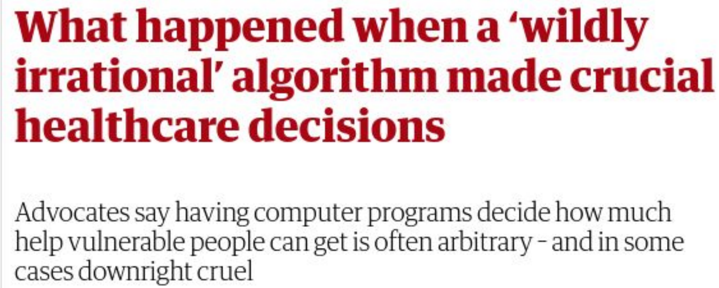


PUBLISHED DECEMBER 5, 2023

HOW AN ALGORITHM DECIDES WHICH FRENCH HOUSEHOLDS TO AUDIT FOR BENEFIT FRAUD

BY MANON ROMAIN, ADRIEN SÉNÉCAT, ELSA DELMAS, LÉA GIRARDOT AND THOMAS STEFFEN

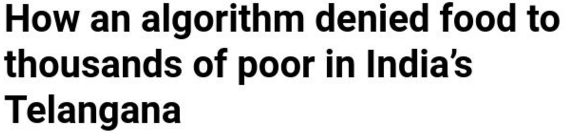
https://www.lemonde.fr/en/les-decodeurs/visuel/2023/12/05/how-an-algorithm-decides-which-french-households-to-audit-for-benefit-fraud_6313254_8.html



What happened when a 'wildly irrational' algorithm made crucial healthcare decisions

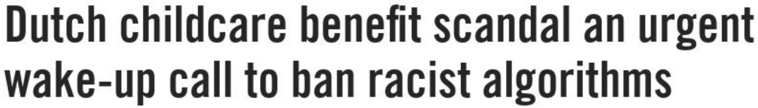
Advocates say having computer programs decide how much help vulnerable people can get is often arbitrary - and in some cases downright cruel

https://www.theguardian.com/us-news/2021/jul/02/algorithm-crucial-healthcare-decisions
2 July, 2021



How an algorithm denied food to thousands of poor in India's Telangana

https://www.aljazeera.com/economy/2024/1/24/how-an-algorithm-denied-food-to-thousands-of-poor-in-indias-telangana Jan 24 2024



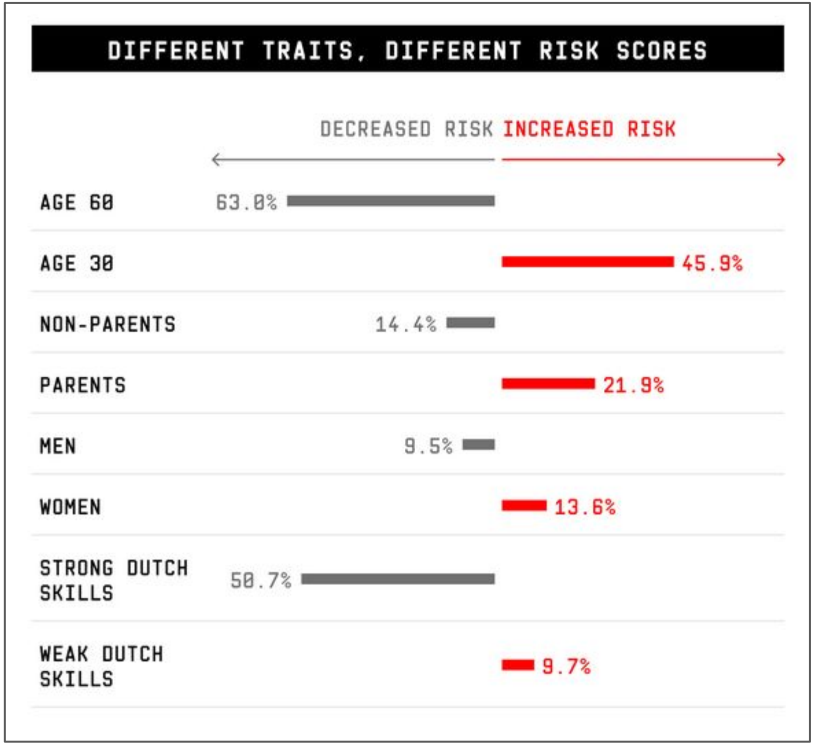
Dutch childcare benefit scandal an urgent wake-up call to ban racist algorithms

https://www.amnesty.org/en/latest/news/2021/10/xenophobic-machines-dutch-child-benefit-scandal/ October 25, 2021

Critical AI Perspective

Discussion

- What are the impacts of using these features for fraud detection?
    - Why do we see the effects shown?
- What were the goals of building this algorithm?
- Could they be different?



Netherlands childcare benefits fraud detection algorithm. Eight out of the 315 features used.
https://www.wired.com/story/welfare-state-algorithms/ 6 March 2023

How can it be different?

Potential source of harm:

- Often introduced to cut costs in (underfunded) benefits systems

Alternative goals:

- Increasing benefit levels
- Simplify enrollment
- Improve working conditions (caseworkers, caregivers)

https://www.lawfaremedia.org/article/the-algorithms-too-few-people-are-talking-about

AI as Socio-technical Systems

- Made by humans, used by humans
- What gets built vs what does not - not neutral
- Complex to untangle the socio-technical interactions
  - Ex: AI and user interaction
    - Disagreeing with algorithmic output?
    - Changing skills in the workplace?

The Abstraction Traps

Selbst et al. (2019) identify five common traps when computer scientists intervene on AI systems (to make fairer, reduce bias and harms)

- Framing Trap
- Portability Trap
- Formalism Trap
- Ripple Effect Trap
- Solutionism Trap

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. Proceedings of the Conference on Fairness, Accountability, and Transparency, 59–68. https://doi.org/10.1145/3287560.3287598

Framing trap

- Assessing the social outcome of a model (e.g. accuracy, fairness) based on a frame that does not incorporate the entire relevant system
    - Ex. a model provides a risk assessment score to a social worker
    - Model is statistically demonstrated not to discriminate based on race or gender
    - But how does the social worker apply the risk score? Is there discretion discriminatory?

Portability trap

- An algorithm for social context will need to incorporate a lot of context
- This means it is unlikely to be suitable for a changed context
- Engineering training, and cited benefits of computation focus heavily on portability

Formalism Trap

- Trying (and failing) to account for complex social concepts mathematically
  - There is rarely universal agreement on definitions
  - Any definition is usually complex and situational
  - E.g. the very nature of concepts like ethical and fair are procedural, contextual, contestable

Ripple Effect Trap

- Failure to account for how the algorithmic technology will change the behaviour of the existing system
    - How does the user respond to / use the algorithmic output?
    - Changes in job skills needed and learned
    - Prioritization of factors quantified in the algorithm, loss of others

Solutionism Trap

- Failure to recognize the possibility that the best solution to a problem may not involve technology

Non-neutrality of AI Technology

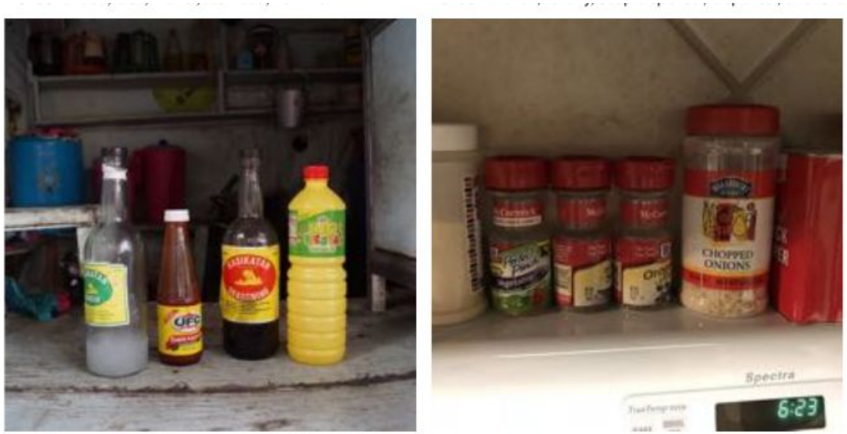
- An AI system’s output is not neutral “a view from somewhere” (Donna Harraway) (Elish & boyd 2018)
- All design decisions are normative decisions
- Example, labeling data (‘ground truth’ creation):

> *“...the work of annotators is profoundly informed by the interests, values, and priorities of other actors above their station.[…] Assigning meaning to data is often presented as a technical matter. This paper shows it is, in fact, an exercise of power with multiple implications for individuals and society.”* (Miceli et al. 2020)

Miceli, M., Schuessler, M., & Yang, T. (2020). Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW2), 1–25. https://doi.org/10.1145/3415186

Elish, M. C., & boyd, danah. (2018). Situating methods in the magic of Big Data and AI. Communication Monographs, 85(1), 57–80. https://doi.org/10.1080/03637751.2017.1375130

Whose View?



Ground truth: Spices **Phillipines, 262 $/month**

Azure: bottle, beer, counter, drink, open
Clarifai: container, food, bottle, drink, stock
Google: product, yellow, drink, bottle, plastic bottle
Amazon: beverage, beer, alcohol, drink, bottle
Watson: food, larder food supply, pantry, condiment, food seasoning
Tencent: condiment, sauce, flavorer, catsup, hot sauce

Ground truth: Spices **USA, 4559 $/month**

Azure: bottle, wall, counter, food
Clarifai: container, food, can, medicine, stock
Google: seasoning, seasoned salt, ingredient, spice, spice rack
Amazon: shelf, tin, pantry, furniture, aluminium
Watson: tin, food, pantry, paint, can
Tencent: spice rack, chili sauce, condiment, canned food, rack

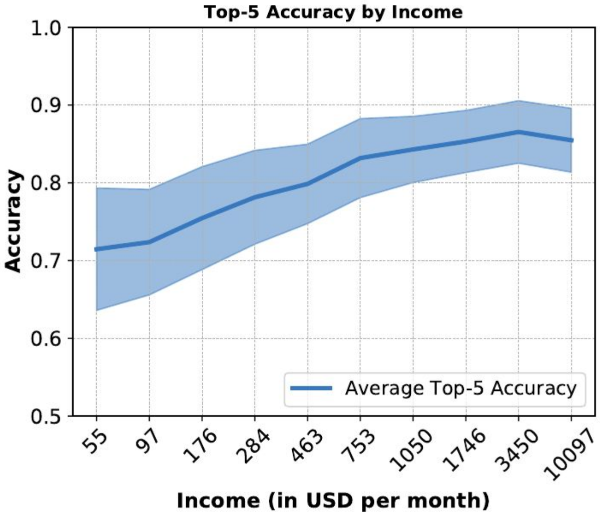
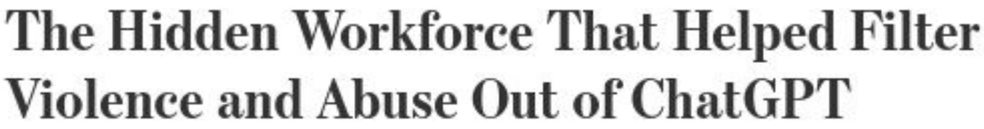


Figure 3: Average accuracy (and standard deviation) of six object-recognition systems as a function of the normalized consumption income of the household in which the image was collected (in US$ per month).

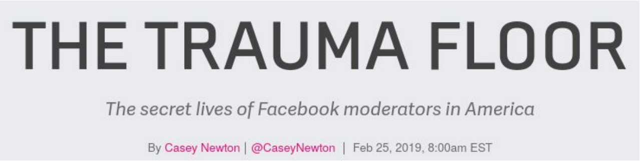
de Vries, T., Misra, I., Wang, C., & van der Maaten, L. (2019). Does Object Recognition Work for Everyone? 52–59.
https://openaccess.thecvf.com/content_CVPRW_2019/html/cv4gc/de_Vries_Does_Object_Recognition_Work_for_Everyone_CVPRW_2019_paper.html
Ex. from: http://cs231n.stanford.edu/slides/2020/lecture_14.pdf

Labour conditions and Material Realities



The Hidden Workforce That Helped Filter Violence and Abuse Out of ChatGPT

Podcast:https://www.wsj.com/podcasts/the-journal/the-hidden-workforce-that-helped-filter-violence-and-abuse-out-of-chatgpt/ffc2427f-bdd8-47b7-9a4b-27e7267cf413 July 11 2023

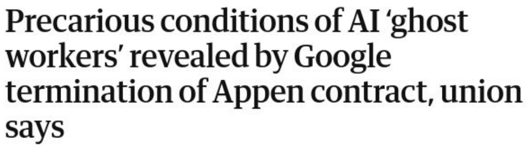


THE TRAUMA FLOOR

The secret lives of Facebook moderators in America

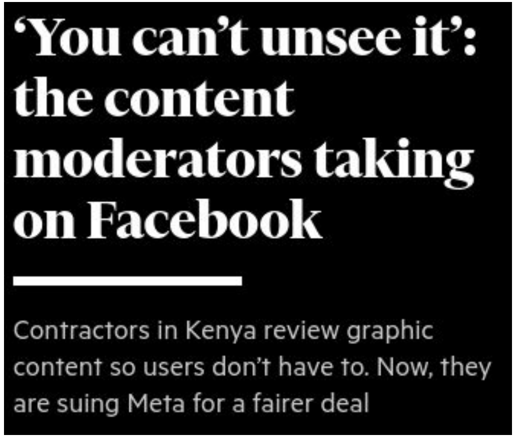
By Casey Newton | @CaseyNewton | Feb 25, 2019, 8:00am EST

https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona February 25 2019



Precarious conditions of AI 'ghost workers' revealed by Google termination of Appen contract, union says

https://www.theguardian.com/australia-news/2024/jan/23/precarious-conditions-of-ai-ghost-workers-revealed-by-google-termination-of-appen-contract-union-says 23 January 2024



'You can't unsee it': the content moderators taking on Facebook

Contractors in Kenya review graphic content so users don't have to. Now, they are suing Meta for a fairer deal

https://www.ft.com/content/afeb56f2-9ba5-4103-890d-91291aea4caa May 18 2023

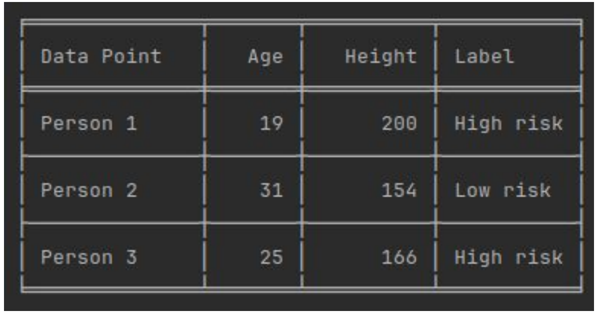
Recommended quick read: https://just-tech.ssrc.org/articles/data-work-and-its-layers-of-invisibility/

Impact of Generative AI

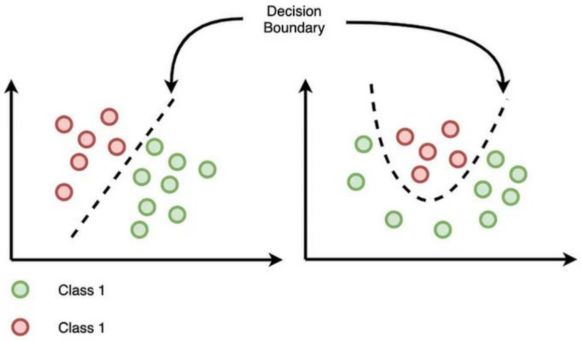
Generative vs. Discriminative Models

Discriminative

- Needs labeled data for training
- Learns the relationship between features and labels
  - the decision boundary
- Good for categorizing data points
- Examples: logistic regression, decisions trees
- Risks: discriminative decisions, allocational harms



Data Point	Age	Height	Label
Person 1	19	200	High risk
Person 2	31	154	Low risk
Person 3	25	166	High risk

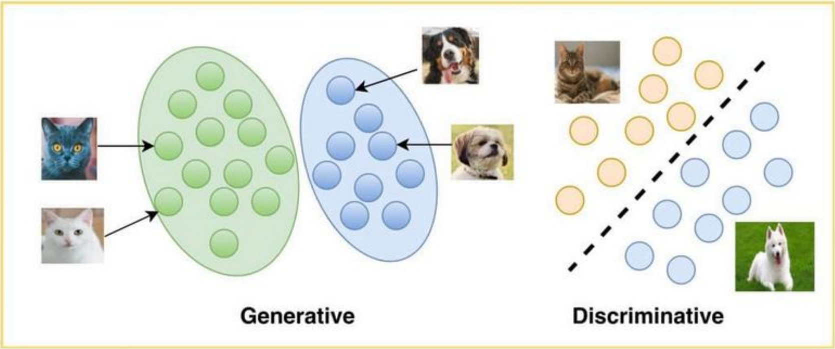


Decision Boundaries

Generative vs. Discriminative Models

Generative

- Learns the underlying patterns in the data (what makes up a cat / dog)
- Can be used to generate new (synthetic) datapoints
- Needs large amount of training data
- Examples: GANS, GPT models
- Can be used for categorization, but may not be as good



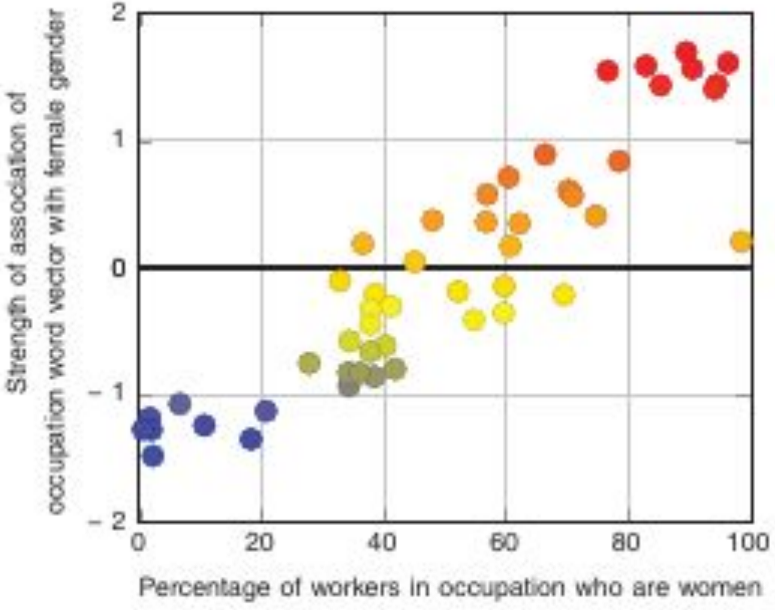
https://learnopencv.com/generative-and-discriminative-models/

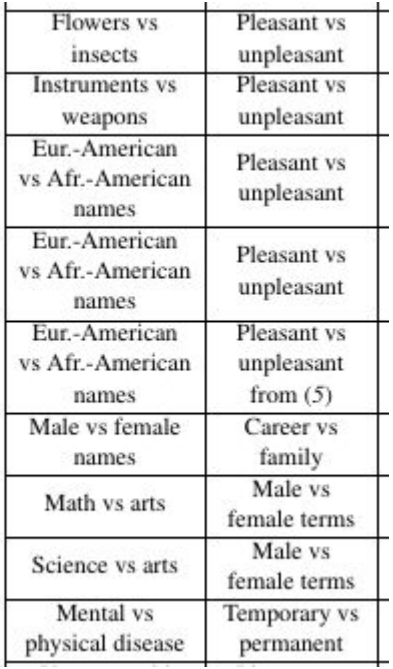
Generative AI - What is learned?

- The languages, customs, views of the training data
- Risks of representational harm
  - Stereotyping - propagates negative generalizations about a social group
  - Incorrect, or no representation of specific groups
- Allocational harm still possible as well

Stereotyped Associations (Caliskan et al. 2017)

- Embeddings: computational representations of words
- Similarity between embeddings = learned association between words
- Many stereotyped associations found





Flowers vs insects	Pleasant vs unpleasant
Instruments vs weapons	Pleasant vs unpleasant
Eur.-American vs Afr.-American names	Pleasant vs unpleasant
Eur.-American vs Afr.-American names	Pleasant vs unpleasant
Eur.-American vs Afr.-American names	Pleasant vs unpleasant from (5)
Male vs female names	Career vs family
Math vs arts	Male vs female terms
Science vs arts	Male vs female terms
Mental vs physical disease	Temporary vs permanent

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183–186. https://doi.org/10.1126/science.aal4230

Generative AI - Environmental Impact

Emissions

- Training GPT-3 - carbon emissions of 502 metric tons of carbon
  - Equivalent of 112 cars over one year
- Daily carbon footprint of 50 pounds of CO2 (8.4 tons per year)
- Impact is still a fraction of the impact of major data centers overall (which is large and increasing)

Li, P., Yang, J., Islam, M. A., & Ren, S. (2023). *Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models* (arXiv:2304.03271). arXiv. http://arxiv.org/abs/2304.03271

https://news.climate.columbia.edu/2023/06/09/ais-growing-carbon-footprint/

Generative AI - Environmental Impact

GPT-3 and Water

- Training GPT-3 consumed an estimated 5.4 million liters of water
  - 700,000 liters on-site for cooling
  - The rest is electricity generation and parts manufacturing

    scope-1 on-site water consumption. Additionally, GPT-3 needs to “drink” (i.e., con-
- A 500ml bottle of water for roughly 10-50 responses from chat GPT-3
  - When and from where has an impacts

Li, P., Yang, J., Islam, M. A., & Ren, S. (2023). *Making AI Less “Thirsty”: Uncovering and Addressing the Secret Water Footprint of AI Models* (arXiv:2304.03271). arXiv. http://arxiv.org/abs/2304.03271

https://news.climate.columbia.edu/2023/06/09/ais-growing-carbon-footprint/

Generative AI - Further Ethical Considerations

- Benefit of who? English language dominance
- Automating false of malicious content
- Source material: copyright and privacy infringement?
- Source material: dangerous information

> *“...while some language technology is genuinely designed to benefit marginalized communities… most language technology is built to serve the needs of those who already have the most privilege in society.”*

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜 Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623. https://doi.org/10.1145/3442188.3445922

Technical Approaches

Introduction

Fair ML is found at the intersection of the **abstract notion of fairness, legislation, and mathematical sciences**.

It is used to detect, assess, and mitigate harmful disparities related to **protected attributes** that may be present in datasets and models through the incorporation of ‘**fairness indicators’**.

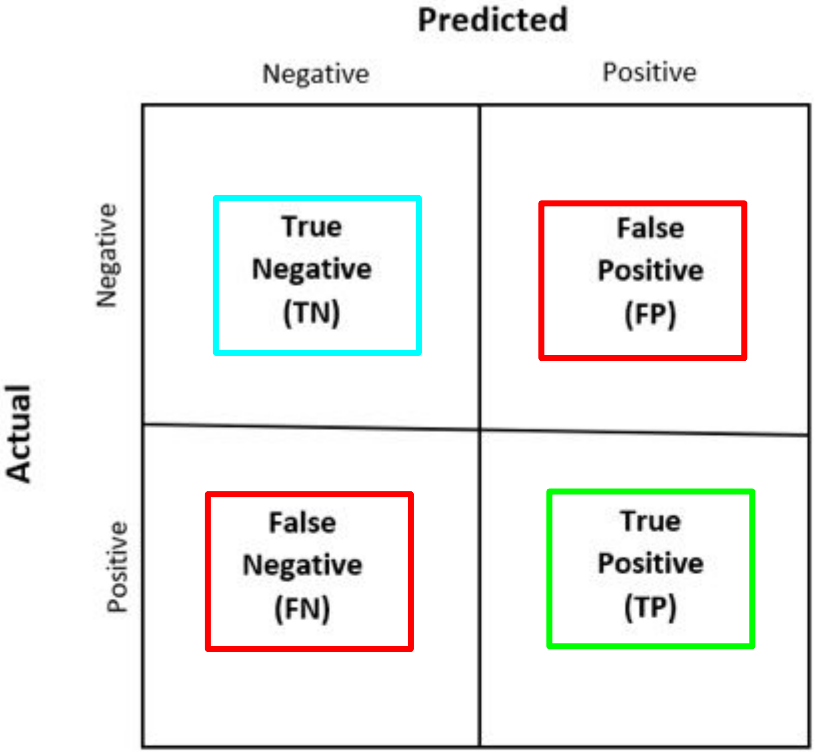
With more than **20 definitions** said to be documented, fairness in machine learning can’t be constrained to a single universal definition.

A. Narayanan. “21 fairness definitions and their politics,” in FAT*,New York, USA., Feb. 2018.

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A Survey on Bias and Fairness in Machine Learning,” arXiv:1908.09635 [cs], Sep. 2019, Accessed: Oct. 08, 2020. [Online]. Available: http://arxiv.org/abs/1908.09635.

The Confusion Matrix

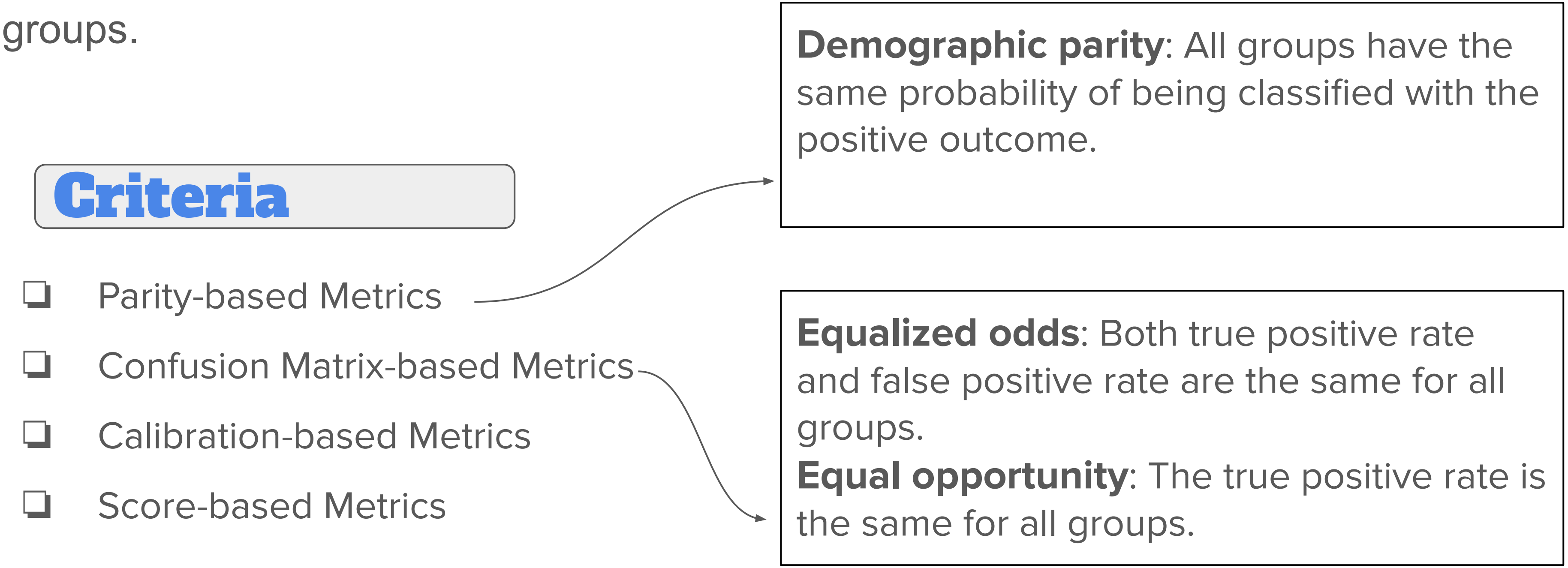
- A measure of model performance
- Requires labels (‘Actual’)
- Loan approval example:
    - Positive = approval
    - Negative = rejection
    - Approval is the ‘Positive’ outcome



		Predicted	
		Negative	Positive
Actual	Negative	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (TP)

Group Fairness

Group-based fairness metrics compare the outcome of the algorithm for two or more groups.



Criteria

- □ Parity-based Metrics
- □ Confusion Matrix-based Metrics
- □ Calibration-based Metrics
- □ Score-based Metrics

Demographic parity: All groups have the same probability of being classified with the positive outcome.

Equalized odds: Both true positive rate and false positive rate are the same for all groups.
Equal opportunity: The true positive rate is the same for all groups.

Individual Fairness

Individual fairness requires that individuals who are similar are treated similarly.

Criteria

❑ Fairness through unawareness

Unawareness: Do not include the protected attributes in the decision.
Fairness: Assumes decisions will be fair if the protected attribute is not taken into account.
Limitation: In some applications (ex. medicine) the protected attribute is needed. Risk of proxy variables.

❑ Fairness through awareness

Awareness: the distance between two individuals based on some distance metric.
Fairness: Individuals who are close are also close in probability of being given the same decision.
Limitation: Choosing a distance function.

Causal Fairness

Utilizes causal graphs to determine fairness.

Criteria

❑ Counterfactual fairness

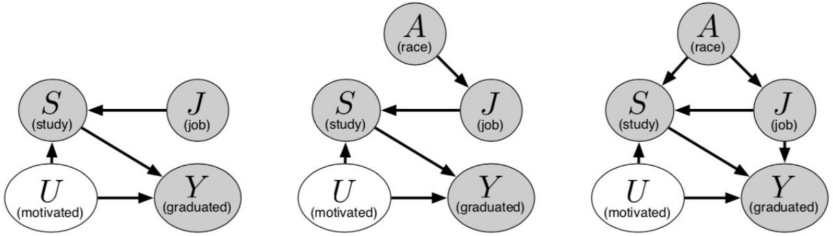


Fig6: some possible causal graphs

A decision is fair towards an individual if it is demonstrated through a causal graph that it would be the same if that individual belonged to a different protected group.

Limitation: Construction of these graphs is not trivial and the treatment of a feature like ‘race’ as a causal element may be at odds with lived experience.[1]

[1] L. Hu and I. Kohler-Hausmann, “What’s Sex Got to Do With Fair Machine Learning?,”

Image from https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb

Challenges

- Mathematical incompatibility of some fairness constraints (you cannot achieve all at once).
- These definitions do not consider the impact on groups or individuals over time.
- Assuming that fairness can be mathematically defined to create a fair system does not address domain-specific societal and historical contexts.
- Do they legitimize inequalities under a justification of “merit"?[1]

[1] Kasy, Maximilian, and Rediet Abebe. “Fairness, Equality, and Power in Algorithmic Decision-Making.” In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 576–86. FAccT ’21. New York, NY, USA: Association for Computing Machinery, 2021. https://doi.org/10.1145/3442188.3445919.

Activity - Breakout Rooms in Jitsi

- Design a ‘fairness metric’ for one of your own projects
- A measurable assessment of a potentially harmful system behaviour
  - When is it measured and how?
  - How is it communicated?
  - How is it intended to be used / reacted to?

Non Technical Approaches - Design Principles

Overarching Design Principles

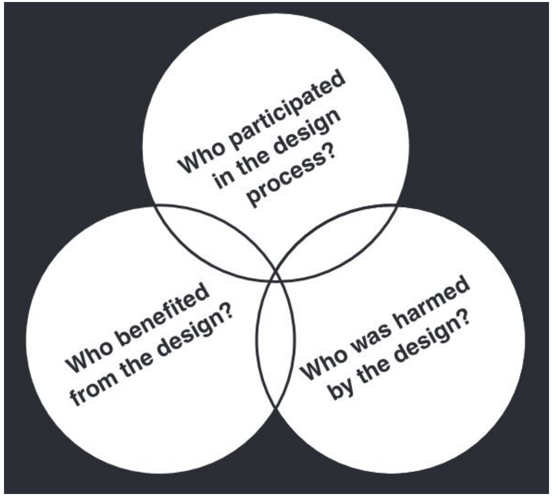
- Include affected people from the beginning
- Recognize / decide on the values behind what is being built
- Acknowledge the normativity of technical decisions
- Assess beyond the model - larger impacts
- Consider how it could be different

Value Sensitive Design

- Account for human values throughout the design process
  - Intrinsic values of the goal
  - Values of direct and indirect stakeholders
- Iterative investigations: conceptual, empirical and technical
  - Conceptual: implicated stakeholders, identifying and defining values, how to engage with tradeoffs
  - Empirical: observation, interviews, experiments etc.
  - Technical: testing technology in relation to values AND proactive design

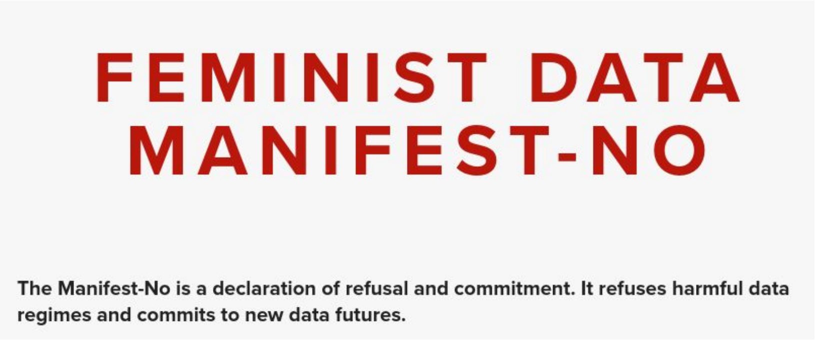
Friedman, B., Kahn, P. H., Borning, A., & Huldtgren, A. (2013). Value Sensitive Design and Information Systems. In N. Doorn, D. Schuurbiers, I. van de Poel, & M. E. Gorman (Eds.), Early engagement and new technologies: Opening up the laboratory (pp. 55–95). Springer Netherlands. https://doi.org/10.1007/978-94-007-7844-3_4
Specific to computational modeling, inspired by VSD: Fish, B., & Stark, L. (2021). Reflexive Design for Fairness and Other Human Values in Formal Models. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 89–99. https://doi.org/10.1145/3461702.3462518

Design Justice (for Social Justice and Social Change)


Who participated in the design process?
Who benefited from the design?
Who was harmed by the design?

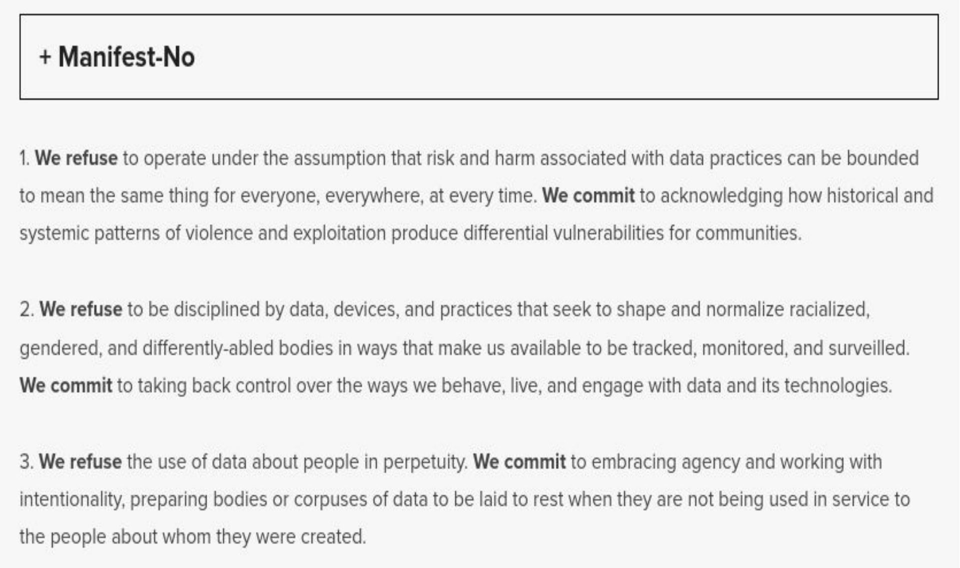
Costanza-Chock, S. (2020). Design Justice: Community-Led Practices to Build the Worlds We Need. The MIT Press.
https://library.oapen.org/handle/20.500.12657/43542 (Book)
https://designjustice.org/zines

Saying No?



FEMINIST DATA MANIFEST-NO

The Manifest-No is a declaration of refusal and commitment. It refuses harmful data regimes and commits to new data futures.



+ Manifest-No

1. **We refuse** to operate under the assumption that risk and harm associated with data practices can be bounded to mean the same thing for everyone, everywhere, at every time. **We commit** to acknowledging how historical and systemic patterns of violence and exploitation produce differential vulnerabilities for communities.

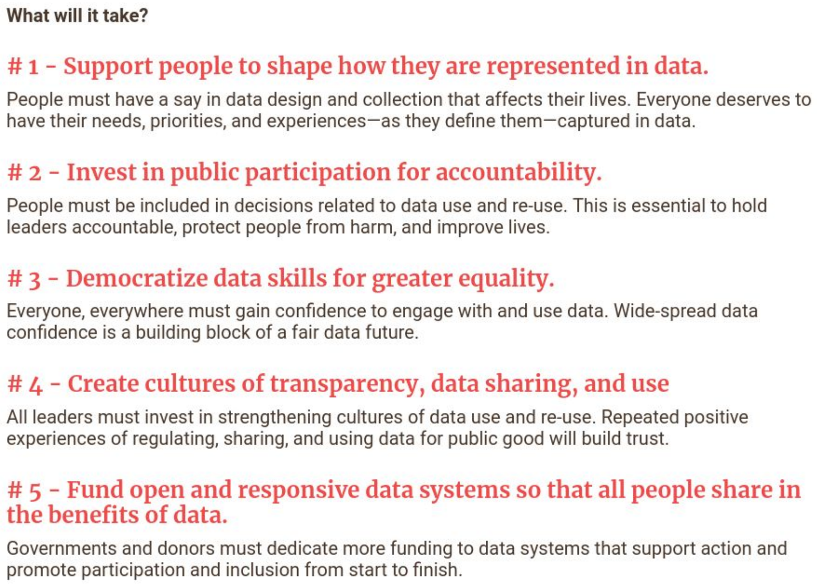
2. **We refuse** to be disciplined by data, devices, and practices that seek to shape and normalize racialized, gendered, and differently-abled bodies in ways that make us available to be tracked, monitored, and surveilled. **We commit** to taking back control over the ways we behave, live, and engage with data and its technologies.

3. **We refuse** the use of data about people in perpetuity. **We commit** to embracing agency and working with intentionality, preparing bodies or corpuses of data to be laid to rest when they are not being used in service to the people about whom they were created.

There are over 30…

Cifor, M., Garcia, P., Cowan, T.L., Rault, J., Sutherland, T., Chan, A., Rode, J., Hoffmann, A.L., Salehi, N., Nakamura, L. (2019). Feminist Data Manifest-No. Retrieved from: https://www.manifestno.com/.

Data Values Manifesto (for Sustainable Development)



What will it take?

1 - Support people to shape how they are represented in data.

People must have a say in data design and collection that affects their lives. Everyone deserves to have their needs, priorities, and experiences—as they define them—captured in data.

2 - Invest in public participation for accountability.

People must be included in decisions related to data use and re-use. This is essential to hold leaders accountable, protect people from harm, and improve lives.

3 - Democratize data skills for greater equality.

Everyone, everywhere must gain confidence to engage with and use data. Wide-spread data confidence is a building block of a fair data future.

4 - Create cultures of transparency, data sharing, and use

All leaders must invest in strengthening cultures of data use and re-use. Repeated positive experiences of regulating, sharing, and using data for public good will build trust.

5 - Fund open and responsive data systems so that all people share in the benefits of data.

Governments and donors must dedicate more funding to data systems that support action and promote participation and inclusion from start to finish.

https://www.data4sdgs.org/datavaluesproject/manifesto-demanding-fair-data-future

More Keywords and Recommended Reads

Participatory Design, co-design

Caselli, T., Cibin, R., Conforti, C., Encinas, E., & Teli, M. (2021). Guiding Principles for Participatory Design-inspired Natural Language Processing. Proceedings of the 1st Workshop on NLP for Positive Impact, 27–35. https://doi.org/10.18653/v1/2021.nlp4posimpact-1.4

Relational Ethics

Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. Patterns, 2(2), 100205. https://doi.org/10.1016/j.patter.2021.100205

Technology for Social Good

Abebe, R., Barocas, S., Kleinberg, J., Levy, K., Raghavan, M., & Robinson, D. G. (2020). Roles for computing in social change. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 252–260. https://doi.org/10.1145/3351095.3372871

Intersectional Feminist Analysis

D’Ignazio, C., & Klein, L. F. (2020). Data feminism. The MIT Press. (Book)

Reflective Questions

- **Who is involved, what is their context? (Klumbyte et al. 2022)**
  - Positionality?
  - Interests?
  - Problems they want to solove?
- **Whose story is not being told here? What are their perspectives? (Klumbyte et al. 2022)**
  - Can be a creative exercise (critical fabulation, speculation, design fictions)
- **Which values are prioritised? (Friedman et al. 2013)**
  - What tradeoffs are being made?
  - How can they be addressed?
- **Who benefits? (Costanza-Chock, S. 2020)**
- **Who is harmed? (Costanza-Chock, S. 2020)**

Klumbytė, G., Draude, C., & Taylor, A. S. (2022). Critical Tools for Machine Learning: Working with Intersectional Critical Concepts in Machine Learning Systems Design. 2022 ACM Conference on Fairness, Accountability, and Transparency, 1528–1541. https://doi.org/10.1145/3531146.3533207

Contributing to computer science / AI technology

- Report on the biases you do (or do not) find be embedded in the AI tools you use
  - particularly in relation to your topics of agriculture, food, and sustainability
- What kind of AI technologies would you like to see, or would you need, to successfully create your ideal project?

Thank you.
Questions?

HUNGRY ECOCITIES

A S+T+ARTS RESIDENCIES PROJECT

In4Art
guiding curiosity....

FundingBox

Studio
Other
Spaces

BRNO
UNIVERSITY
OF TECHNOLOGY

eat
this.

KU LEUVEN

Mendel
University
in Brno

CARLO
RATTI
ASSOCIATI®

Starts.eu

S+T+ARTS

European
Commission

HungryEcoCities project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement 101069990.