Intelligent Machines, Ethics and Law (COMP2400/6400)

Algorithmic Bias – Lecture

Dr Regina Fabry

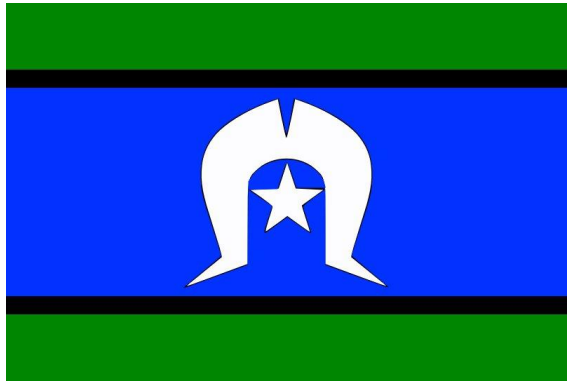
School of Humanities, Discipline of Philosophy

Macquarie University

Week 2 – 3 March 2025

MACQUARIE University
SYDNEY·AUSTRALIA

We acknowledge the Traditional Custodians of the land on which Macquarie University stands – the Wallumattagal Clan of the Dharug Nation – whose cultures and customs have nurtured, and continue to nurture, this land since time immemorial. We pay our respects to the Elders, past and present.



Introduction: 5 Questions

1. What is bias?
2. What is algorithmic bias?
3. What are (some of) the reasons and causes of algorithmic bias?
4. How can algorithmic bias be mitigated?
5. Why might at least some mitigation strategies not be successful?

What Is Human Cognitive-Affective Bias? Two Definitions

- De-contextualised definition: The notion of ‘bias’ refers to a systematic deviation from a moral, statistical, or social standard in situations of assessment and decision-making (Fazelpour & Danks, 2021).
- Contextualised definition: The notion of ‘bias’ refers to the unfair and unjust treatment of members of (intersecting) structurally oppressed groups in situations of assessment and decision-making.
- In both cases, biases can be explicit (deliberate and accessible to awareness) or implicit (nondeliberate and not accessible to awareness).

What Is Human Cognitive-Affective Bias?
The Contextualised Definition

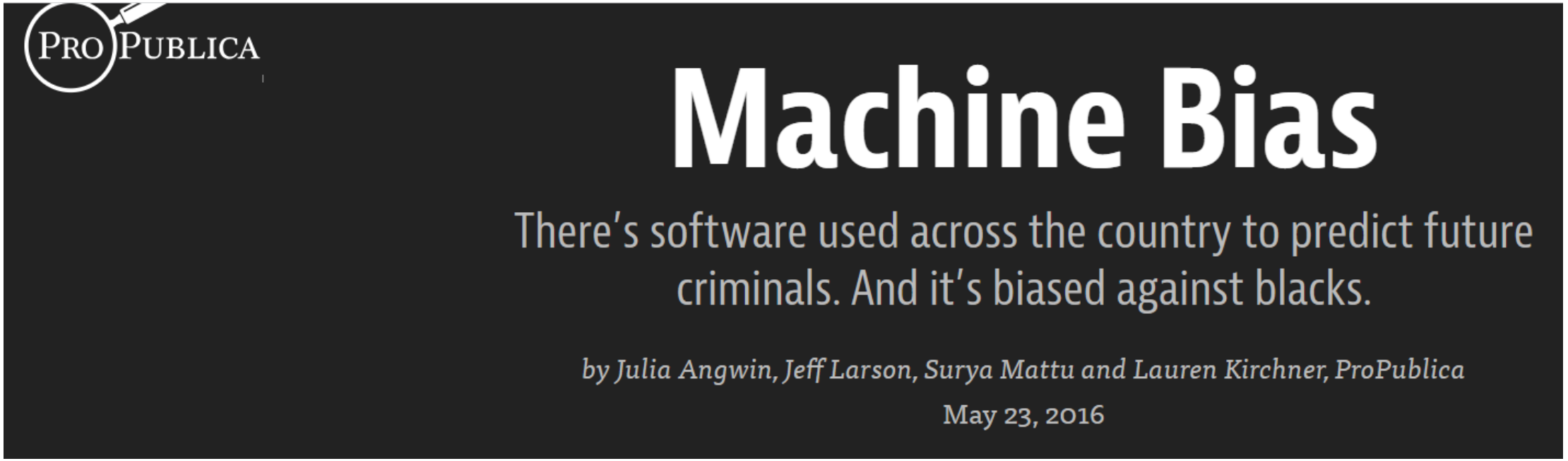
Contextualised definition: The notion of ‘bias’ refers to the unfair and unjust treatment of members of (intersecting) structurally oppressed groups in situations of assessment and decision-making.

- ‘Oppression’ is a structural, multi-faceted notion that captures “[…] the vast and deep injustices some groups suffer as a consequence of often unconscious assumptions and reactions of well-meaning people in ordinary interactions, media and cultural stereotypes, and structural features of bureaucratic hierarchies and market mechanisms […].” (Young, 1990, p. 41)
- Relevant, often intersecting forms of structural oppression include sexism, misogyny, homophobia, transphobia, racism, xenophobia, ableism, classism, and agism.

What Is Algorithmic Bias?

- De-contextualised definition: “[…] algorithmic bias is simply systemic deviation in algorithmic output, performance, or impact, relative to some norm or standard […]. An algorithm can be morally, statistically, or socially biased (or other), depending on the normative standard used.” (Fazelpour & Danks, 2021, p. 2)
- Contextualised definition: The notion of ‘bias’ refers to the unfair and unjust treatment of members of (intersecting) structurally oppressed groups as a result of algorithmic decision-making (pace Johnson 2021).
- Most research on algorithmic decision making has operated with, or tacitly assumed, some version of the contextualised definition of ‘algorithmic bias.’

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)
An Example of Algorithmic Bias



https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)
An Example of Algorithmic Bias

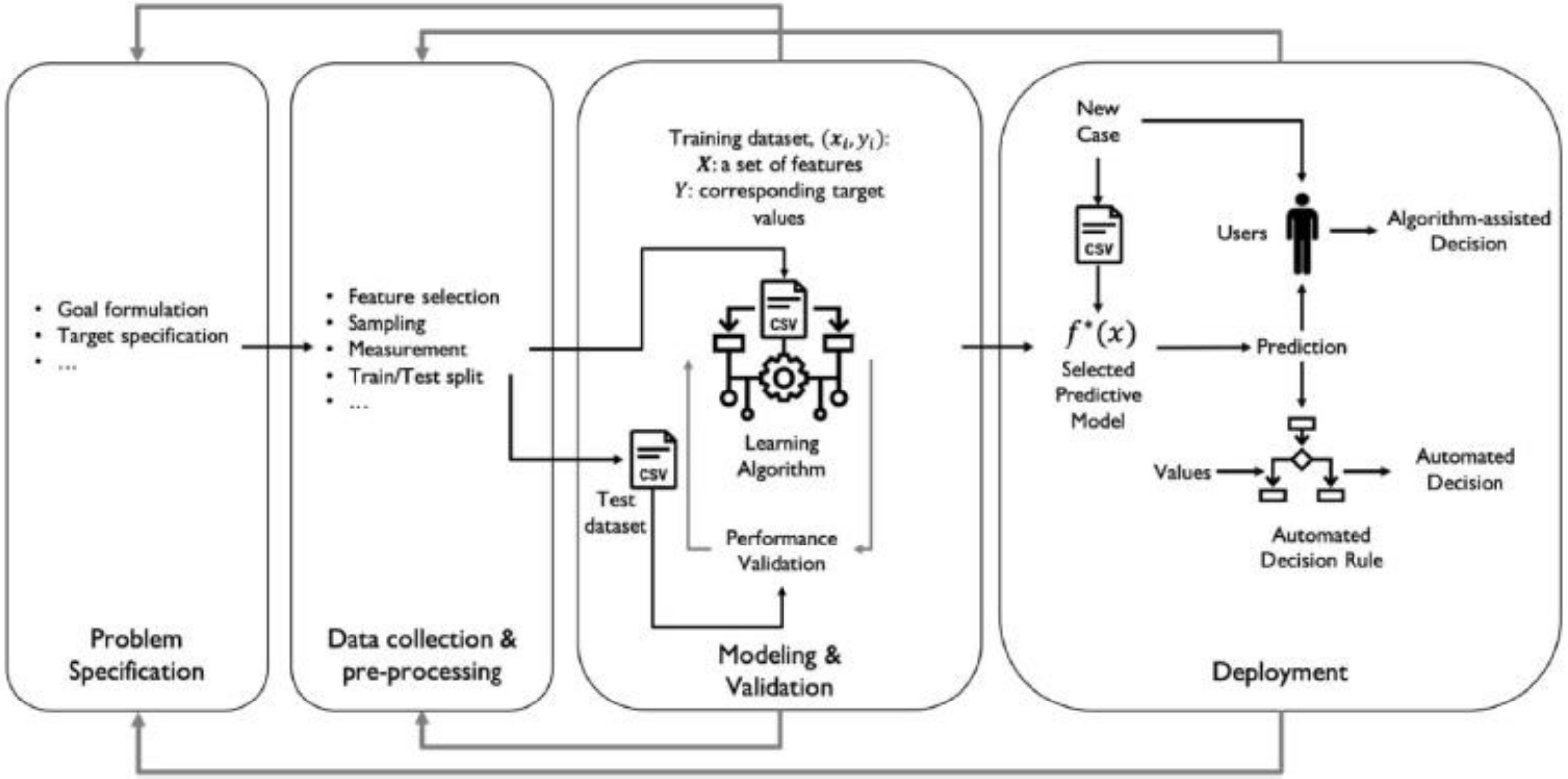
Algorithmic Bias and the Value Neutrality Thesis

- According to the *value neutrality thesis*, “[…] technological artifacts such as algorithms do not embody or implement values, and so their *use* is the only appropriate locus of ethical evaluation.” (Fazelpour & Danks, 2021, p. 3; italics in original)
- Value neutrality, and thus the absence of bias, is unfeasible for at least two reasons:
    1. “[…] algorithms make possible certain kinds of decisions and capabilities, and so embody the value that those decisions or capabilities are important.” (Ibid.)
    2. “[…] algorithms implement values because they are almost always optimized for performance relative to a standard” (Ibid.). This standard is in itself a value judgment.

Algorithmic Bias and the Value Neutrality Thesis Cont'd

- Work on algorithmic bias shares assumptions and insights with work in other areas of philosophical research.
- Recent research on situated cognition and affectivity suggests that many resources and practices in the socio-cultural environment are not value neutral.
- They manifest and perpetuate forms of oppression, including things (Liao & Huebner, 2021) and technologies (Spurrett, 2024).
- For example, oppressive things can be defined as "[…] material artifacts and spatial environments that are in congruence with an oppressive system." (Liao & Huebner, 2021, p 94)
- If decision-making is not value neutral, then at what stages can biases become an issue?

Algorithmic Decision-Making
From Problem Specification to Deployment

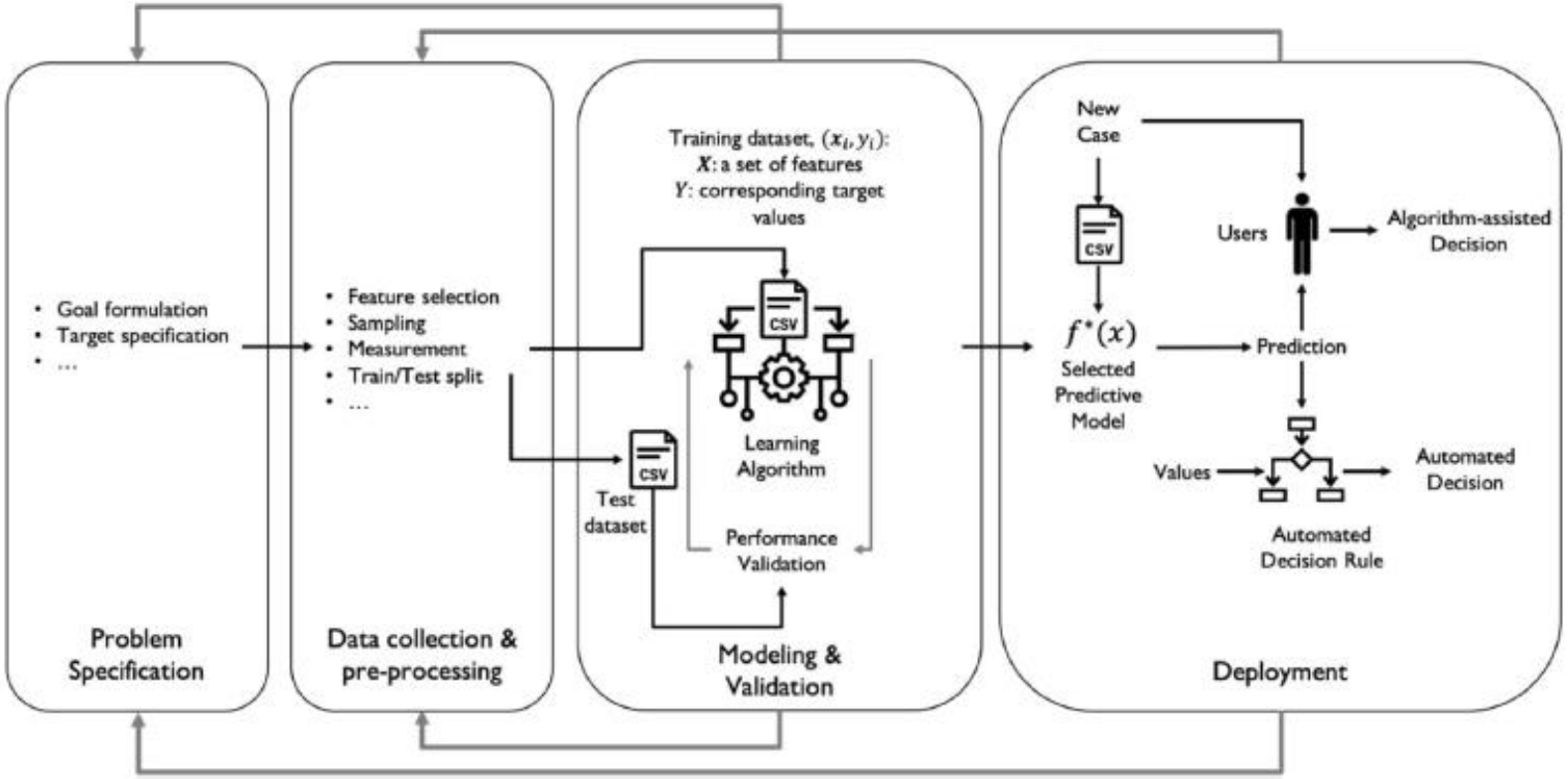


Fazelpour & Danks (2021), Figure 1.

Sources of Algorithmic Bias
Problem Specification

- What exactly is the decision-making problem that needs to be solved?
- Problem specification "[…] requires thinking about our overall aims, the actions available to us, and ways of using the algorithm to help achieve those aims […]." (Fazelpour & Danks, 2021, p. 4)
- What are the values and standards that contribute to the specification of the decision-making problem at issue?
- Which target variables associated with persons and groups are most likely to be affected, intentionally or unintentionally, by algorithmic decision-making?
- Problem specification can lead to *disparate impacts*, whereby "[…] members of a protected group are differentially impacted relative to a (more) dominant group […]." (Ibid.)

Algorithmic Decision-Making
From Problem Specification to Deployment



Fazelpour & Danks (2021), Figure 1.

Sources of Algorithmic Bias
Training Data

- By definition, machine learning (ML) algorithms "[…] output models that partially 'mirror' the statistics in the historical data […]." (Fazelpour & Danks, 2021, p. 6)
- Biases in the training data will be reflected in the outcomes of algorithmic decision-making. These can occur for at least two, often co-occurring reasons:
    1. They can "[…] result from existing biases in real-world systems that are measured in the data." (Ibid.; see also Johnson, 2021)
    2. They can "[…] arise from limitations and biases in our measurement methods" (e.g., non-representative training data). (Ibid.)

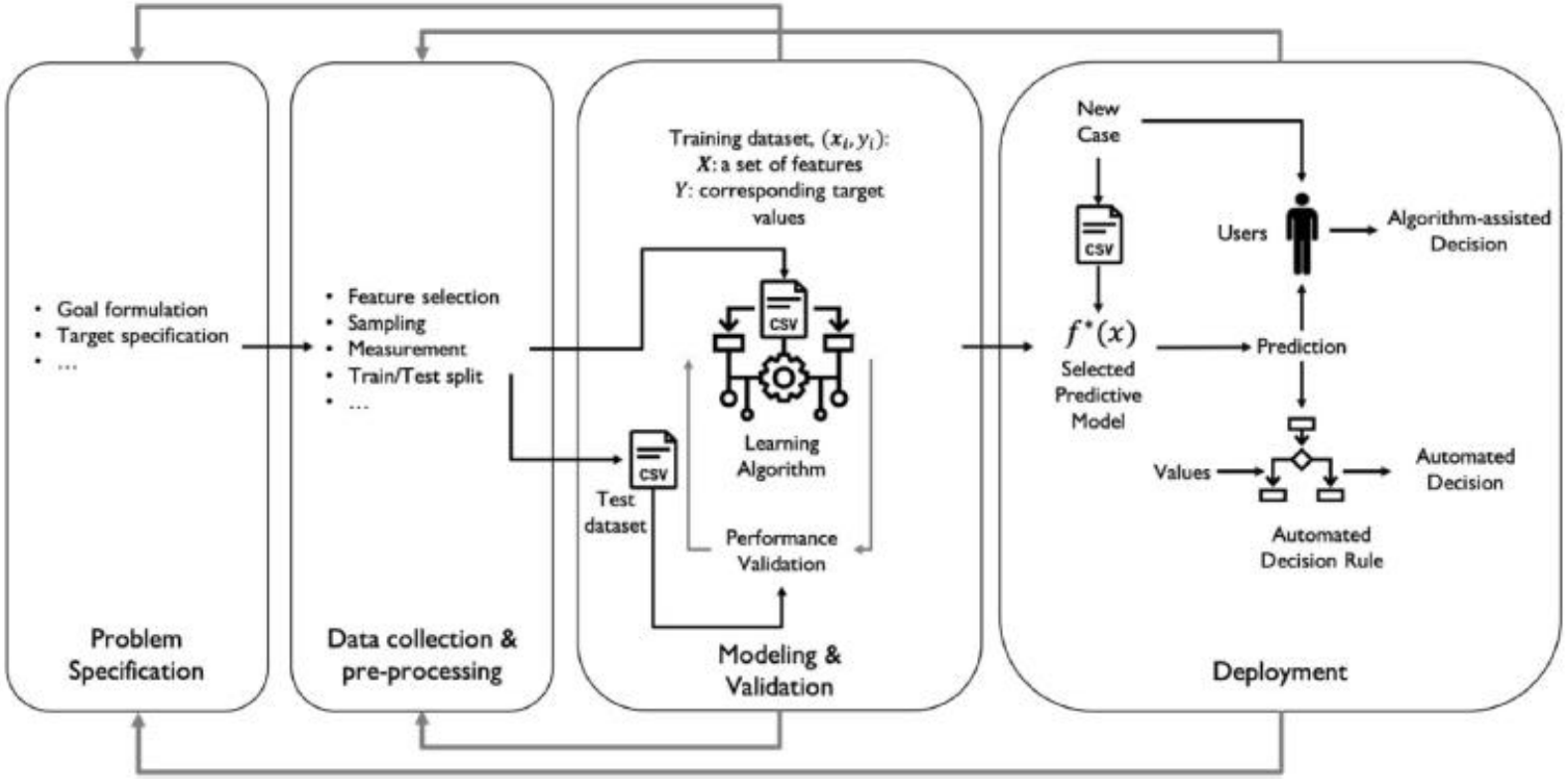
Sources of Algorithmic Bias
Training Data: The Problem of *Ground Lies*

- Ideally, ML algorithms would represent and predict the so-called *ground truth*, patterns in the real world in an accurate and truthful way.
- Many training data are not representative of complex, socially shaped real-world phenomena across a wide range of domains (e.g., health care, employment, immigration, law enforcement) (see Johnson, 2021).
- Rather than reaching ground truths, algorithmic decision-making might lead to algorithmic biases that produce new *ground lies* (Bender, 2024).
- “If we do not actively work to curate the data sets that we want, then we will be collecting data sets that are representative of dehumanizing ideologies such as white supremacy and calling lies ‘ground truth’ […].” (Bender, 2024, p. 116)

Sources of Algorithmic Bias
Training Data: The *Proxy Problem*

- Even if ground lies were to be avoided, algorithmic bias can be enabled by the *proxy problem* (Johnson, 2021).
- The proxy problem consists in the perpetuation of oppressive biases through the collection and pre-processing of *proxy attributes.*
- They are defined as “[…] seemingly innocuous attributes that correlate with socially sensitive attributes, serving as proxies for the socially-sensitive attributes themselves.” (Ibid., p. 9952)
- Proxy attributes in the training data contribute to the perpetuation of biases against members of structurally oppressed groups.

Algorithmic Decision-Making
From Problem Specification to Deployment

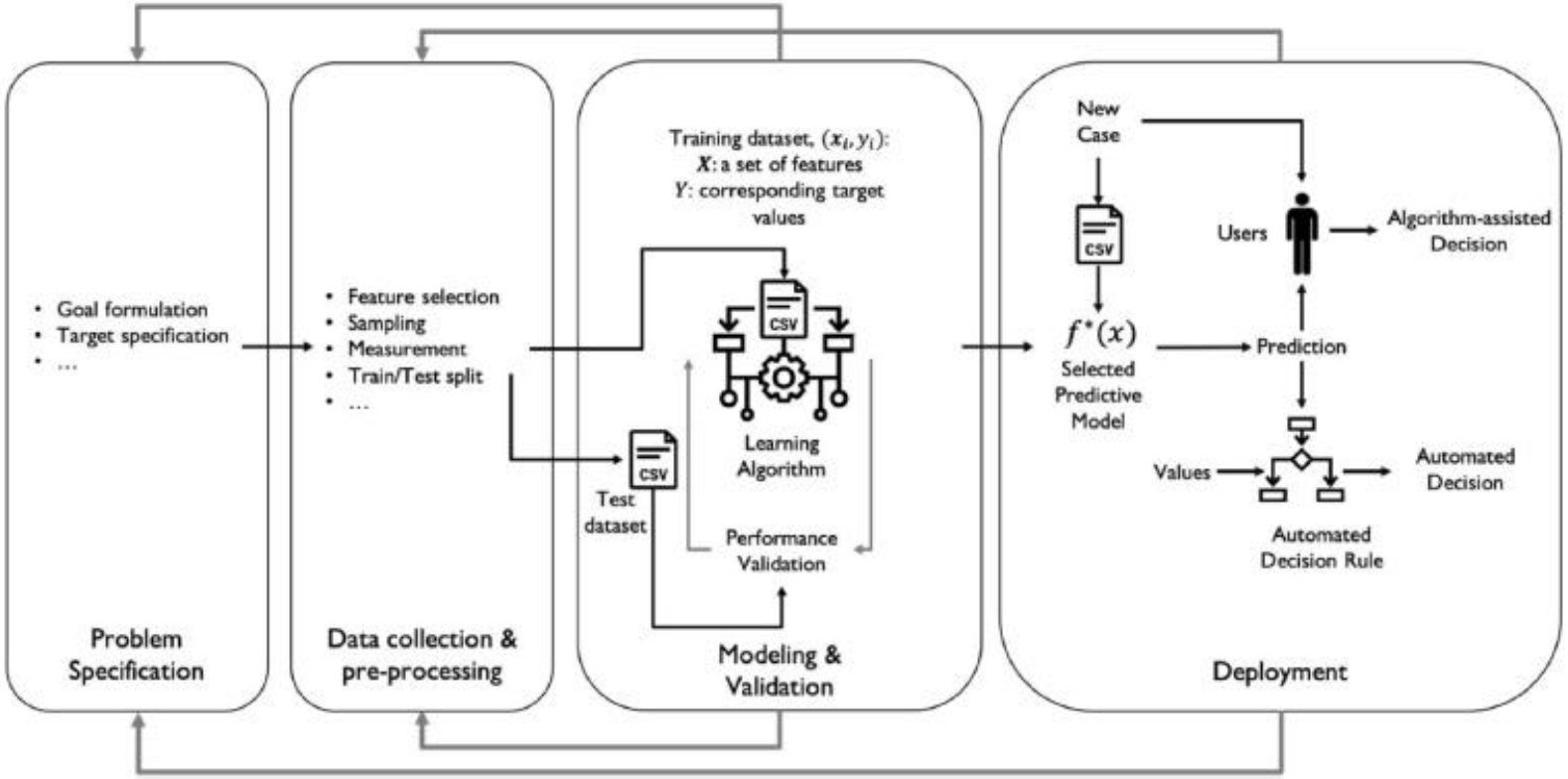


Fazelpour & Danks (2021), Figure 1.

Sources of Algorithmic Bias
Modelling and Validation

- Biases at the stage of problem specification re-occur at the stage of modelling and validation.
- What are the success criteria of algorithmic decision-making, given the specification of a target domain, goals, values, and standards (see Fazelpour & Danks, 2021)?
- What values and norms do different stakeholders (e.g., developers, end-users) request to be reflected in the decision-making algorithms (see Ibid.)?
- Model validation and optimisation might come at the cost of exacerbating *ground lies* and the *proxy problem*.

Algorithmic Decision-Making
From Problem Specification to Deployment

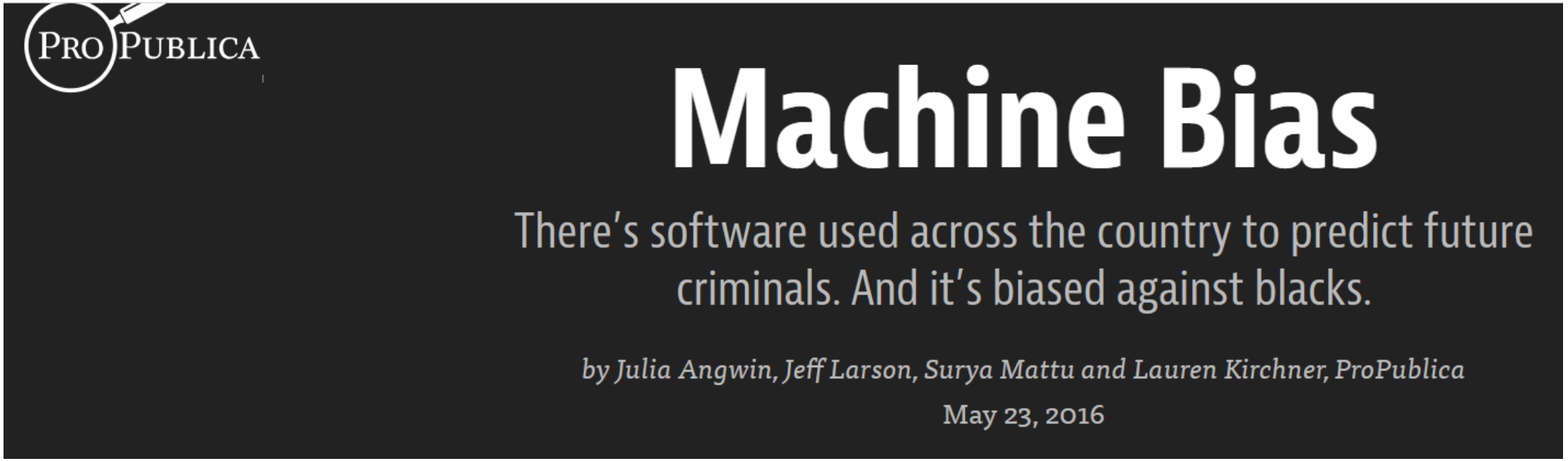


Fazelpour & Danks (2021), Figure 1.

Sources of Algorithmic Bias
Deployment

- Algorithmic decision-making can perpetuate or exacerbate existing biases in high-stakes cases of deployment (e.g., recidivism risk assessment).
- Biases occurring at the stages of problem specification, training data selection and pre-processing, and modelling and validation influence decision-making outcomes in specific cases of deployment.
- The question is not only *what* is employed, but also *how* it is employed by human agents.
- “[…] morally problematic decisions and unjust harms result from a biased algorithm supporting an unbiased human, an unbiased algorithm supporting a biased human, or both being biased.” (Fazelpour & Danks, 2021, p. 8)

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)
An Example of Algorithmic Bias Revisited



https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Mitigating Algorithmic Bias
Fairness through Unawareness

- The *fairness through unanwareness* position entertains the hope that "[…] an algorithm could not possibly be biased about property X [e.g., gender, racialized identity] if it is never told about whether an individual possesses X." (Fazelpour & Danks, pp. 8-9)
- This strategy fails because of the proxy problem (Johnson, 2021): "Other variables in the input data will often be correlated with, and so serve as proxies for, protected attributes; even when these sensitive attributes are not explicitly included in the data, they are 'redundantly encoded' in these proxies […]." (Ibid., p. 9)

Mitigating Algorithmic Bias
Fairness through Unawareness Cont'd

- Attempts to avoid or resolve the proxy problem appears to come at the cost of reducing the accuracy or predictive power of algorithmic decision-making.
- "Thus, a dilemma arises: due to patterns of oppression being to deeply engrained in our social environment, the more we eliminate reliance on proxy attributes in decision-making, the more likely it is that we'll have an accurate and ineffective decision-making procedure." (Johnson, 2021, p. 9957)

Gabbrielle Johnson

Mitigating Algorithmic Bias
Fairness through Mathematical Fairness Measures

1. **Individual measures**: Mathematical representation of ‘treating similar cases similarly.’
   However, it is unclear how similarities of social properties can be quantified.
2. **Statistical measures**: Mathematical representation of “[…] statistical disparities (e.g., accuracy, or error rates relating to sensitivity or specificity) between algorithmic predictions across different protected attributes that we typically think morally ought not impact decisions.” (Fazelpour & Danks, 2021, p. 9)
   However, these measures are at times incompatible with each other and might not be able to do justice to intersectional social attributes.

Mitigating Algorithmic Bias
Fairness through Mathematical Fairness Measures Cont'd

3. **Causal or counterfactual measures**: Mathematical representation of "[…] the causal reasons for patterns of injustice, including what would have occurred (counterfactually) if the biases had not been present […]." (Fazelpour & Danks, 2021, p. 10)

   However, injustices are the result of complex historically entrenched, temporally extended structural processes of domination and oppression.

   Mapping and mathematically representing the causal dynamics underlying biases leading to the perpetuation of these injustices is not straightforward.

Summary

- According to the contextualised definition, the notion of ‘algorithmic bias’ captures the unfair and unjust treatment of members of (intersecting) structurally oppressed groups as a result of algorithmic decision-making (pace Johnson 2021).
- Algorithms are never value neutral, but are situated in complex social structures of domination and oppression.
- The sources of algorithmic biases can be found at various stages: problem specification, collection and pre-processing of training data, modelling and validation, and deployment by end users (Fazelpour & Danks, 2021).
- Due to the social complexities of domination and oppression, injustice and unfairness, the development and application of feasible mitigation strategies has proven to be challenging.

And Next…
… AI and Epistemic Injustice

Philosophical Studies (2025) 182:185–203
https://doi.org/10.1007/s11098-023-02095-2

Check for updates

Algorithmic profiling as a source of hermeneutical injustice

Silvia Milano[1] · Carina Prunkl[2]

Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES 2024)

Epistemic Injustice in Generative AI

Jackie Kay[1 3], Atoosa Kasirzadeh[2 4], Shakir Mohamed[1]
[1]Google DeepMind
[2]Google Research
[3]University College London
[4]University of Edinburgh
kayj@google.com, atoosa.kasirzadeh@gmail.com, shakir@google.com

References

- Bender, E. M. (2024). Resisting dehumanization in the age of “AI.” *Current Directions in Psychological Science*, *33*(2), 114–120. https://doi.org/10.1177/09637214231217286
- Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass*, *16*(8), e12760. https://doi.org/10.1111/phc3.12760
- Johnson, G. M. (2021). Algorithmic bias: On the implicit biases of social technology. *Synthese*, *198*(10), 9941–9961. https://doi.org/10.1007/s11229-020-02696-y
- Liao, S., & Huebner, B. (2021). Oppressive things. *Philosophy and Phenomenological Research*, *103*(1), 92–113. https://doi.org/10.1111/phpr.12701
- Spurrett, D. (2024). On hostile and oppressive affective technologies. *Topoi*. https://doi.org/10.1007/s11245-023-09962-x
- Young, I. M. (1990). Five faces of oppression. In *Justice and the politics of difference* (pp. 39–65). Princeton University Press.