The ethics of algorithms: key problems and solutions.

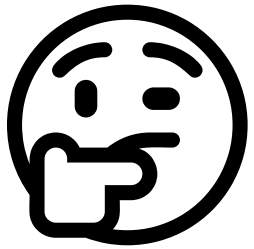
Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., & Floridi, L. (2021). *AI & SOCIETY*, 1-16.

Algorithms

- Algorithms: defined as 'mathematical constructs, their implementations as programs and configurations (applications), and the ways in which these can be addressed.' (p. 216)
- Algorithms that are considered in the paper:
  1. Turn data into evidence for a given outcome
  2. Used to trigger and motivate an action that may have ethical consequences
  3. An attribution of the responsibility for the effects of actions that an algorithm may trigger
- Examples: recommendation, translation, search engine, advertisement, etc.

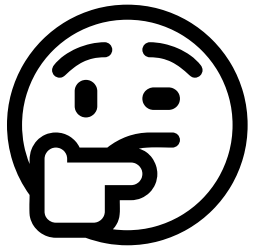
  → Potential to improve individual and social welfare

Algorithms – What can go wrong?



Algorithms – What can go wrong?

recommendation, translation, search engine, advertisement, decision-making (in government, court, financial institutes, schools, hospitals, …), etc.



Algorithms – What can go wrong?

- Fairness
- Accountability
- Interpretability
- Responsibility
- Trustworthiness
- Transparency
- Reliability
- …

Overview

Goal of the paper:

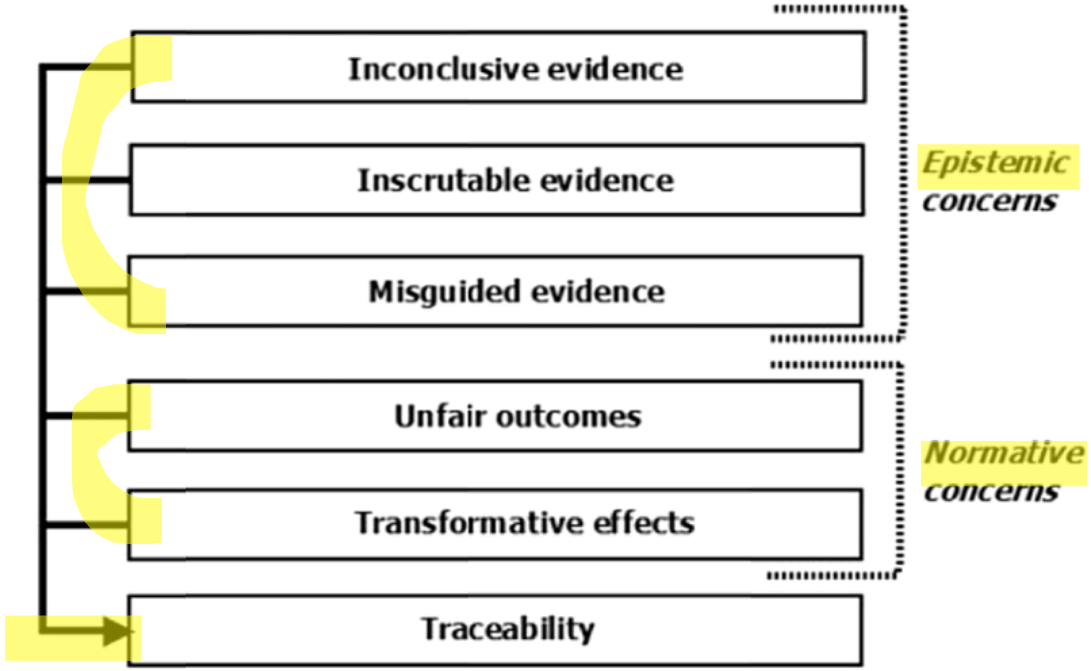
- Contribute to the debate on the identification and analysis of the ethical implications of algorithms
- Provide an analysis of epistemic and normative concerns
- Offer actionable guidance for the governance of the design, development and deployment of algorithms

By providing:

- Systematic search and review on ethics of algorithms

Map of the Ethics of Algorithms

Three types of concerns



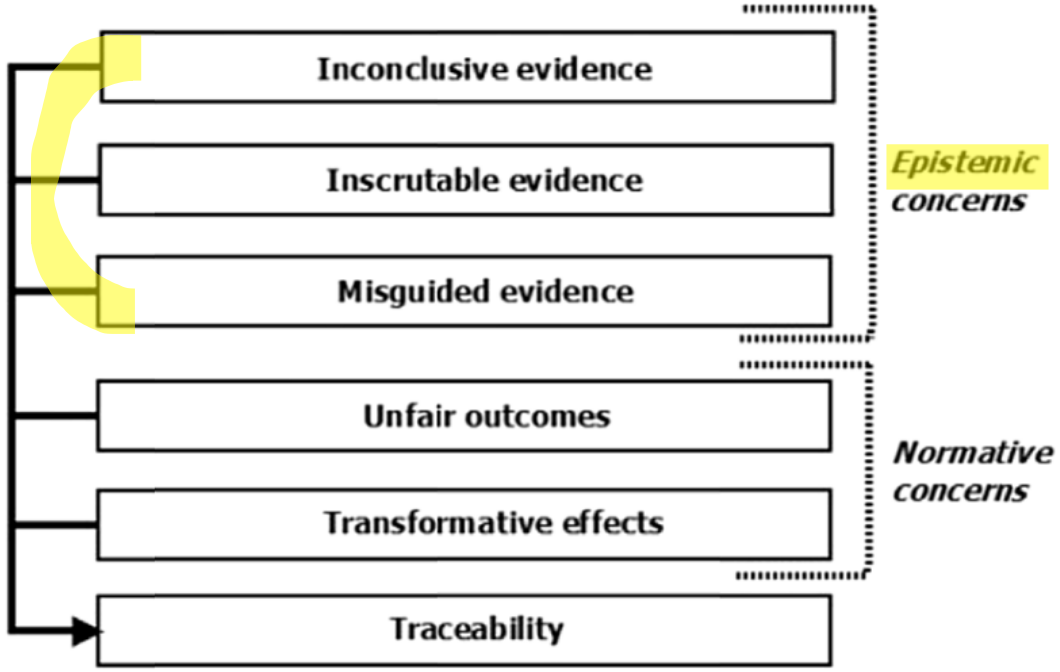
→ knowledge

→ norms, values

→ related to both epistemic and normative concerns

Six types of ethical concerns raised by algorithms

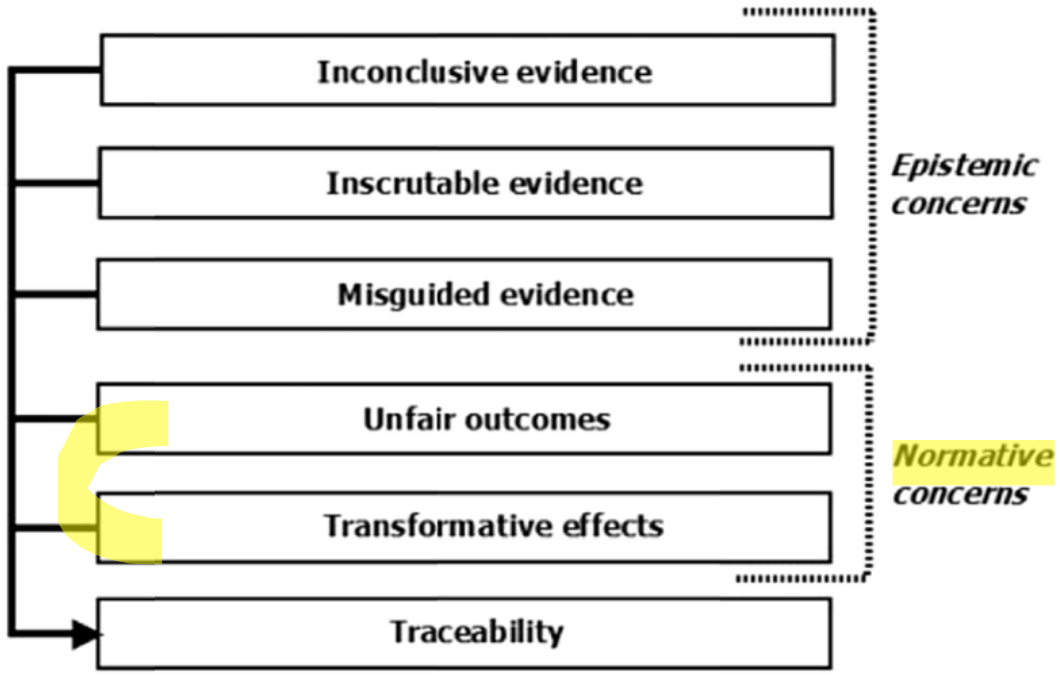
Map of the Ethics of Algorithms



Epistemic concerns

- Highlights the relevance of the quality and accuracy of the data for the justifiability of the algorithm
- What conclusions that algorithms reach?
- How they shape morally-loaded decisions affecting individuals, societies, and the environment?

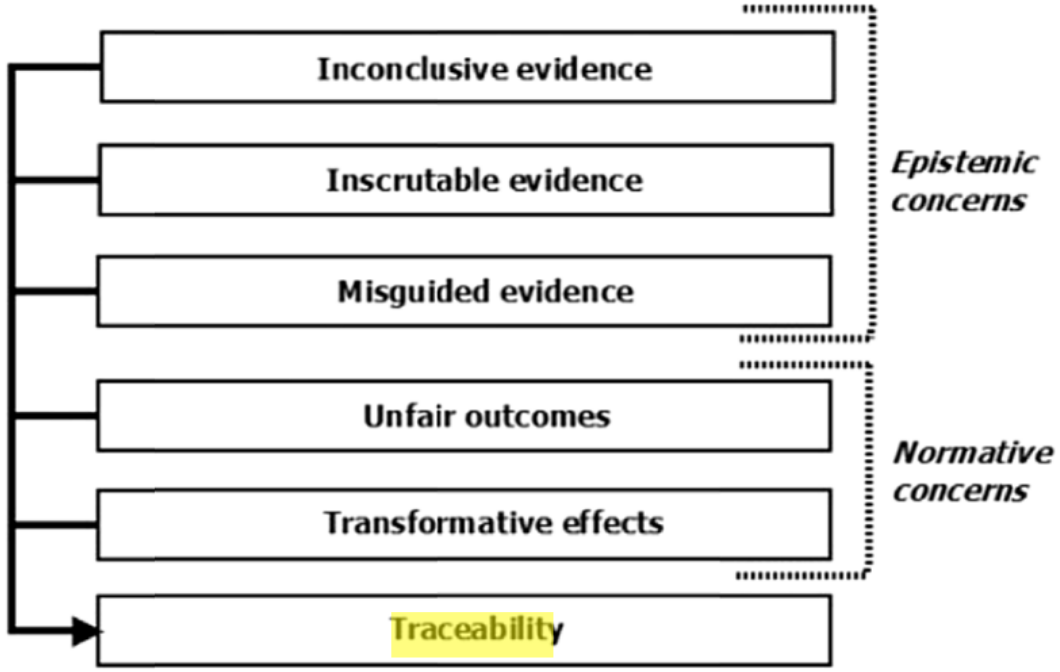
Map of the Ethics of Algorithms



Normative concerns

- Refer ethical impact of algorithmically-driven actions and decisions
- Such as lack of transparency (opacity) of algorithmic processes, unfair outcomes, and unintended consequences.

Map of the Ethics of Algorithms

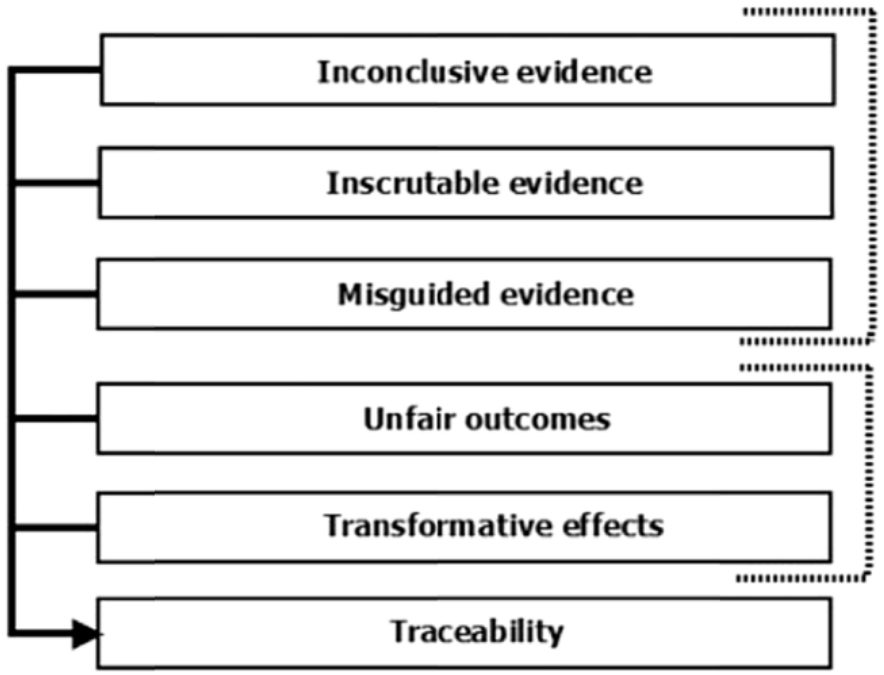


Traceability

- Epistemic and normative concerns *(with the distribution of the design, development and deployment of algorithms)* make it hard to trace the chain of events and factors leading to a given outcome
- It hinders the possibility of identifying its cause and of attributing moral responsibility for it

Map of the Ethics of Algorithms

Why matters?



→ leading to unjustified actions

→ leading to opacity

→ leading to unwanted bias

→ leading to discrimination

→ leading to challenges for autonomy and informational privacy

→ leading to moral responsibility

Fig. 1 Six types of ethical concerns raised by algorithms (Mittelstadt et al. 2016, 4)

1. Inconclusive evidence → unjustified actions

Inconclusive: If research or evidence is inconclusive, it has not proved anything.

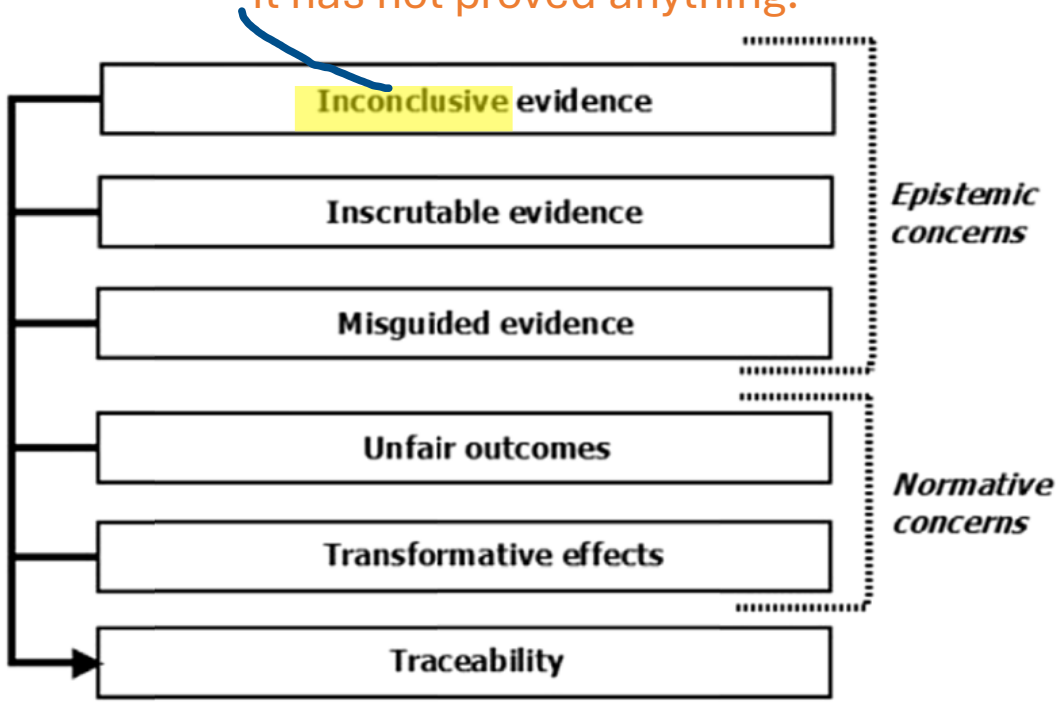


Fig. 1 Six types of ethical concerns raised by algorithms (Mittelstadt et al. 2016, 4)

- ML algorithms produce probabilistic outputs
- Association and correlation between variables, not causal connections
- Can distract attention from the underlying causes of a given problem
- Data quality, assumption guided the data collection rocess, etc. constrains the question that can be answered using a given dataset
- Non-quantifiable inputs (such as willingness to live in clinical settings) are ignored

2. Inscrutable evidence → opacity

If a person or their expression is inscrutable, it is very hard to know what they are really thinking or what they mean.

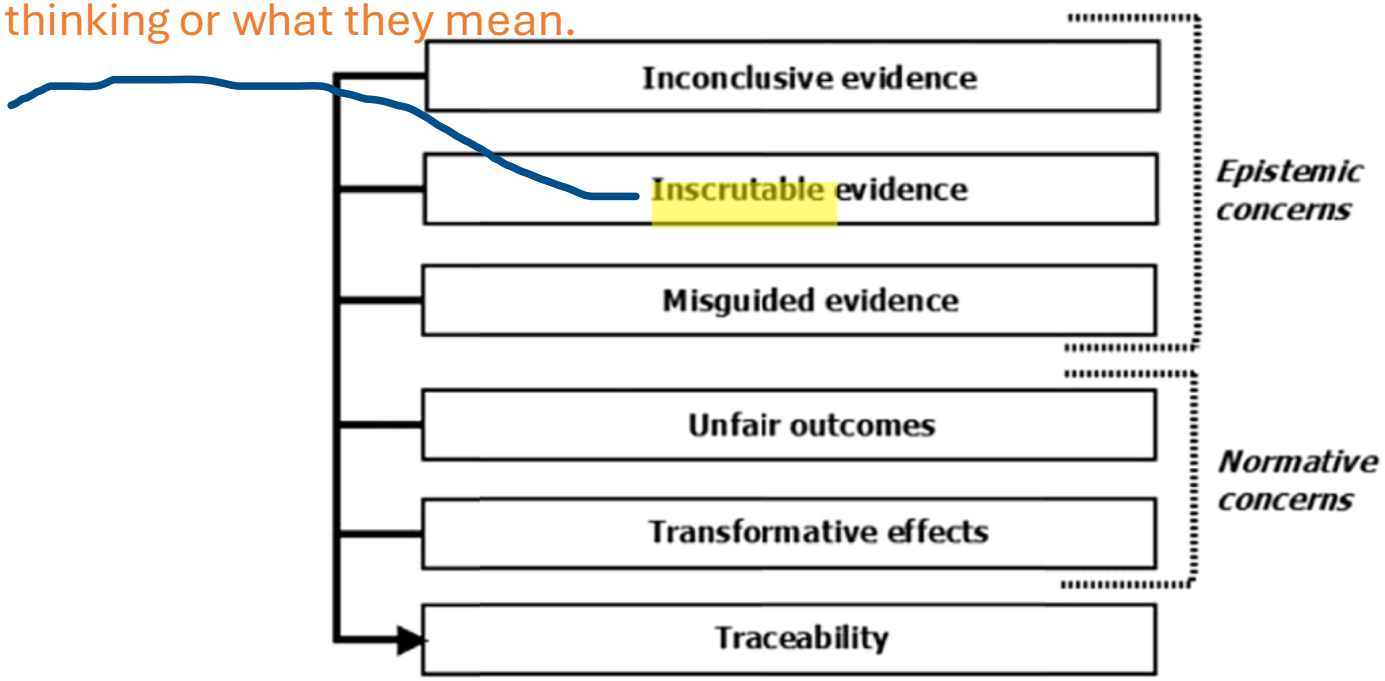


Fig. 1 Six types of ethical concerns raised by algorithms (Mittelstadt et al. 2016, 4)

- Lack transparency that characterises algorithms
- Leads to lack of scrutiniy, accountability, 'trustworthiness'
- Contributing factors:
  - Cognitive impossibility for humans to interpret
  - Lack of appropriate tools to visualize and track large volume of code and data
  - Poorly structured code and data
  - Ongoing updates and human influence over a model
  - Malleability of algorithms
    - reprogrammed in a continuous, distributed, and dynamic way → permanent state of destabilization

3. Misguided evidence → unwanted bias

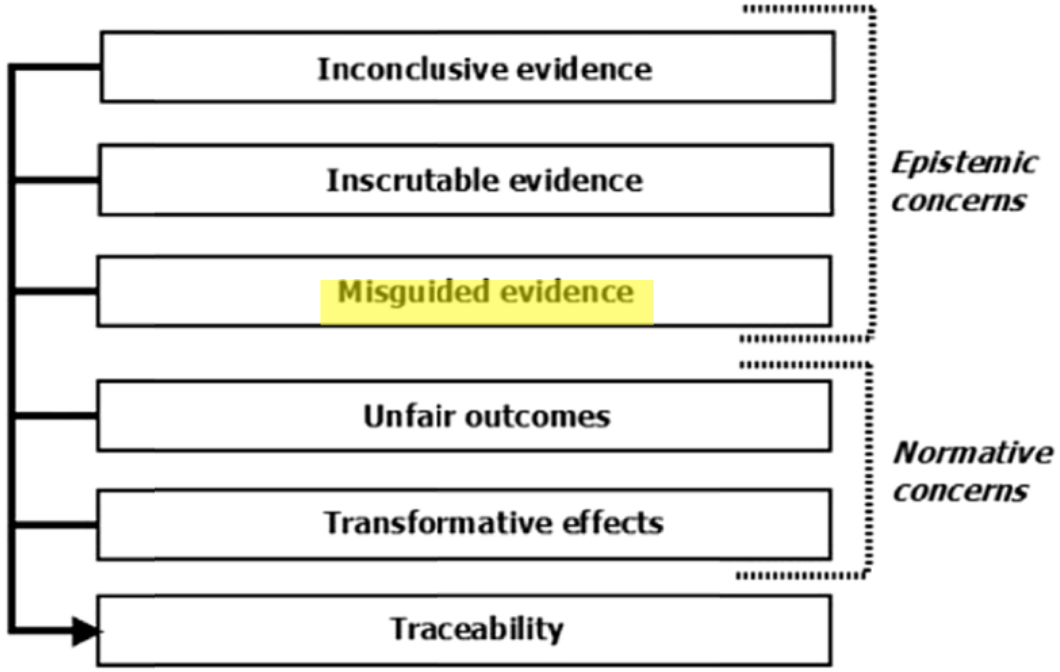


Fig. 1 Six types of ethical concerns raised by algorithms (Mittelstadt et al. 2016, 4)

- "algorithmic formalism" - developer's primary focus on a certain goal
- It tends to ignore social complexity of the real world in the process… with the *illusion of precision*
- Possible abstraction traps fail to account for social context

Five abstraction traps (Selbst et al. 2019)

- A failure to model the entire system over which a social criterion, such as fairness, will be enforced
- A failure to understand how repurposing algorithmic solutions designed for one social context may be misleading, inaccurate, or otherwise do harm when applied to a different context
- A failure to account for the full meaning of social concepts such as fairness, which can be procedural, contextual, and contestable, and cannot be resolved through mathematical formalisms
- A failure to understand how the insertion of technology into a existing social system changes the behaviours and embedded values of pre-existing system;
- A failure to recognize the possibility that the best solution to a problem may not involve technology

4. Unfair outcomes → discrimination

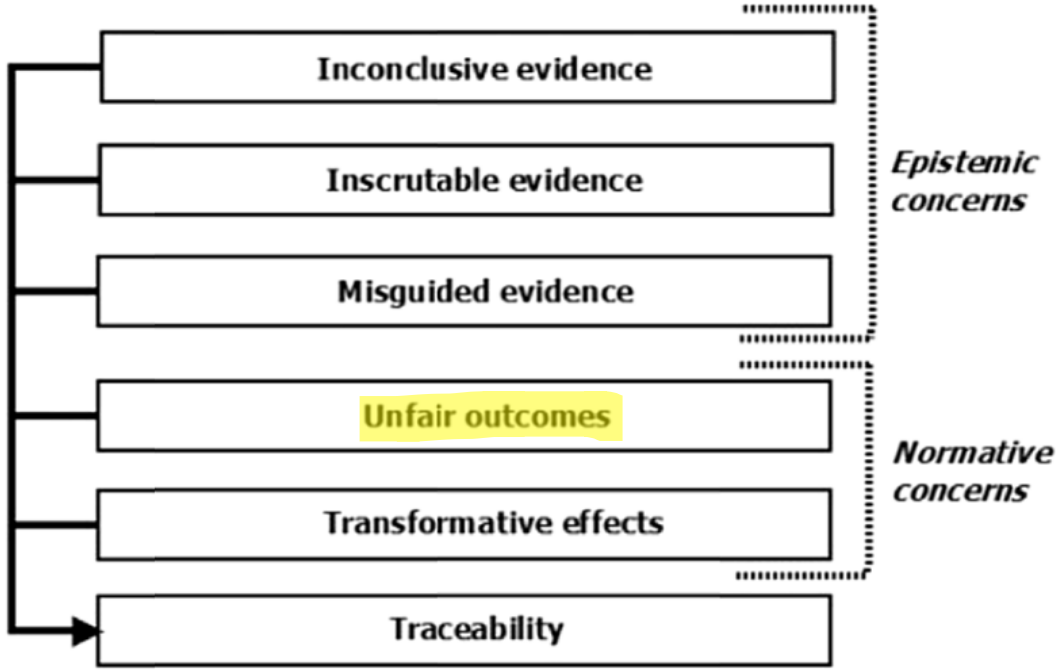


Fig. 1 Six types of ethical concerns raised by algorithms (Mittelstadt et al. 2016, 4)

- Lack of agreement among researchers on the definition, measurements and stadnard of algorithmic fairness
- e.g. four popular definitions
  1. anti-classification: protected categories not being explicitly used in decision-making
  2. classification parity: model being fair of common measures of predictive performance, including false positive and negative rates, are equal across protective groups
  3. calibration: fairness as a measure of how well-calibrated an algorithm is between protected groups
  4. statistical parity: fairness as an equal average probability estimate over all members of protected groups

5. Transformative effects → challenges for autonomy and informational privacy

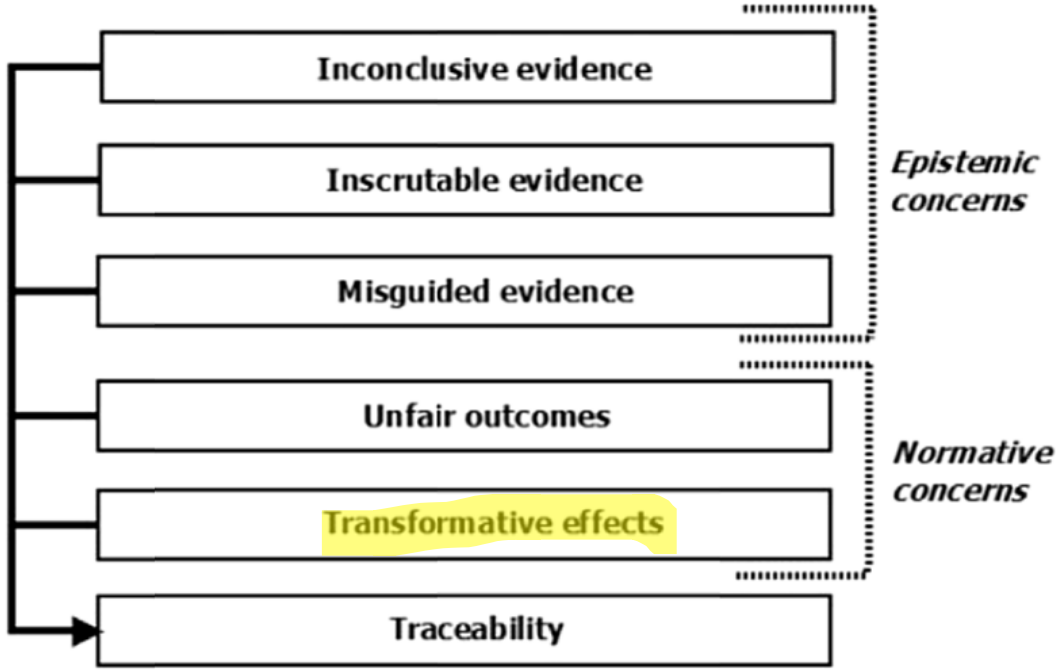


Fig. 1 Six types of ethical concerns raised by algorithms (Mittelstadt et al. 2016, 4)

1. Autonomy: An ecosystem of complex, socio-technical issues can hinder the authonomy of users

- Sources of limiting users' autonomy:
  1. pervasive distribution and proactivity of (learning) algorithms to inform users' choice
  2. Users' limited understanding of algorithms
  3. Lack of second-order power (or appeals) over algorithmic outcomes
  4. informational privacy

2. Informational Privacy: linked with user autonomy

- "guarantees peoples' freedom to think, communicate, and form relationships, among other essential human activities" (Rachels 1975, Allen 2011)
- Increasing interaction with algorithmic systems reduce people's ability to control who has access to information that concerns them and what is being done with it

6. Traceability → moral responsibility

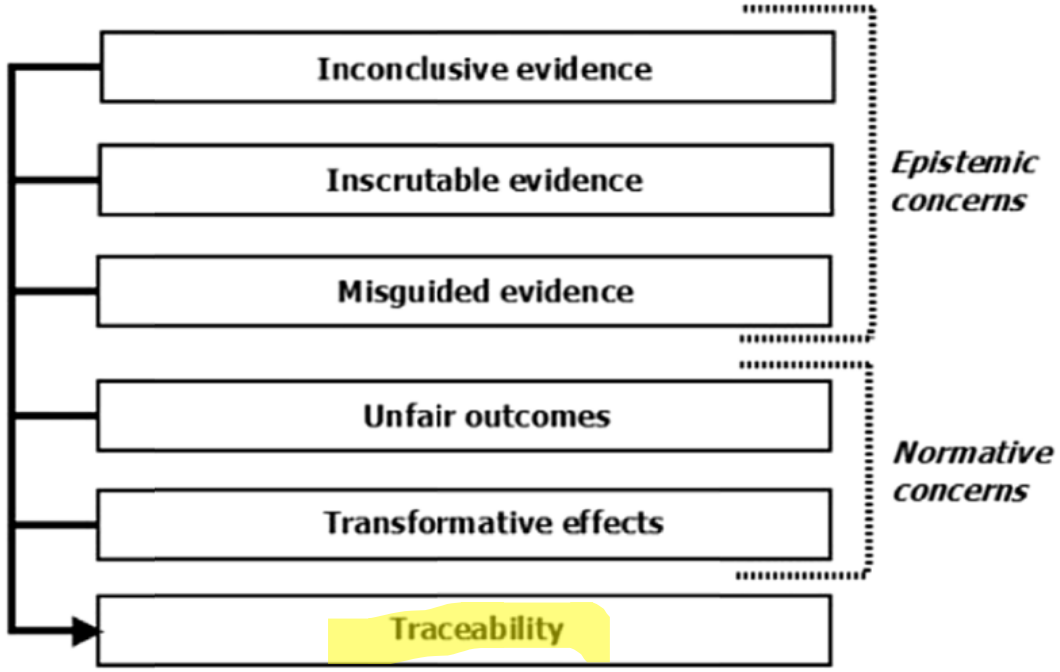


Fig. 1 Six types of ethical concerns raised by algorithms (Mittelstadt et al. 2016, 4)

- "common blurring between technical limitations of algorithms and the broader legal, ethical, and institutional boundaries in which they operate' (Reddy et al. 2019)
- Structure and operation of data brokerage market – impossible to trace the original source of the data in the marketplace
- 'agency laundering': "a moral wrong which consists in distancing oneself from morally suspect actions, regardless of whether those actions were intended or not, by blaming the algorithm" (Rubel et al. 2019)
- Enables avoidance of responsibility due to the interplay between field experts and ML algorithms - '*the computer said no*'

What do you think?

- Is this framework good?
- Or is it missing something?

Thank you!

:)