Ethics of Algorithmic Decision-Making

Simone Collier

University of Toronto

Introduction

The use of algorithms for decision making is becoming more and more prevalent as statistical learning methods improve and data becomes abundant. The ethical considerations for algorithm development and implementation are vast.

We will look at the ethics for alorithms that:

1. Turn data into evidence for a specific outcome.
2. Then use this outcome to make decisions may have ethical consequences.

Ethical Concerns

There are 6 main ethical concerns relating to the use of algorithms.

- Inconclusive evidence
- Inscrutable evidence
- Misguided evidence
- Unfair outcomes
- Transformative effects
- Traceability

Inconclusive Evidence

Inconclusive evidence can lead to unjustified actions.

- The results that are produced by machine learning methods on data always have a level of uncertainty.
- These methods can identify correlations in data but causality is rarely established.
- We see correlations on large sets of data being seen as sufficient evidence for action.
- Even if causality is established this may only apply at the population level while decisions may be being made about individuals.
  (Ex: Insurance premiums are calculated for a population but paid by the individuals.)

With all the uncertainty of our results the ethics of making decisions on the basis of these results begs the question: *what is the consequence of being wrong?*

Inscrutable Evidence

Inscrutable evidence can lead to opacity in the decision making process.

- When data is used as evidence for a conclusion we expect the connection between the two to be comprehensible.
- This is not always the case for machine learning methods.
  - *Accessibility* can be a factor if the algorithms are proprietary.
  - *Comprehensibility* can be a factor if the model has low interpretability (black box).
- Transparency in algorithms make them easier to control, monitor, and correct.
- Failure to render the connection between the data and decision comprehensible
  - disrespects the agency of the data subjects.
  - makes it more difficult to challenge the decisions.

Misguided Evidence

Misguided evidence can lead to bias.

- "Garbage in, garbage out": conclusions are only as reliable as the data they are based on.
- Removing humans from decision making alleges lack of bias, but
  - The design of an algorithm will still reflect the choices and values of its designer (we have seen in this module that there is no one correct choice in model creation).
  - Social biases are sometimes intentional but they can be incorporated unintentionally by the data used to train the algorithm etc.
- The results from the model require interpretation which can reflect the interpreter's biases.

Unfair Outcomes

Unfair outcomes can lead to discrimination in decision making.

- Profiling algorithms identify correlations an make predictions at the group-level.
  - An individual is assessed based on connections to groups rather than actual behaviour.
  - Bias in the algorithm results in discriminatory decision effects.
- Even when attributes such as gender or race are excluded from the data, proxies remain that can influence the algorithm.
- Discriminatory analyses can result in self-fulfilling prophecies and stigmatisation.

Transformative Effects

Transformative effects undermining autonomy.

- Personalisation algorithms are those that are not the same across a sample and are meant to predict the behaviour of an individual.
- They filter information presented to the user based on their preferences and beliefs.
- This should improve decision making by providing only the relevant information but choosing what is relevant is subjective.
- This can influence their decisions; their autonomy can be undermined when their choices are a reflection of the desires of the algorithm's proprietor.
- Reducing the diversity of the information presented can also hinder decision autonomy by not having both sides of the story.

Transformative Effects

Transformative effects leading to challenges for informational privacy.

- Due to the negative outcomes from data sharing such as discrimination and de-individualisation, data subjects want to protect their personal data from third parties.
- There is no one privacy norm that applies to all types of data since the value of the data is inherently linked to the processing.
- Profiling assembles individuals into groups so personal identity is irrelevant.
- This can breach the subject's informational identity since the algorithm uses information about others to describe the subject.
- Regulations in many places do not recognize this as a breach of privacy since the person is not identifiable within the data set.

Traceability

Traceability leading to moral responsibility.

- When an algorithm fails ethically who should be held responsible?
  - In non-learning algorithms the authors should be responsible since they designed the decision rules.
  - The designer of machine learning algorithms cannot assume full responsibility since they do not have enough control over the machine's actions.
- Accountability gap.

References

Mittelstadt, B. D., et al. 2016. The ethics of algorithms: mapping the debate. Big Data & Society.