Fragile Algorithms and Fallible Decision-Makers: Lessons from the Justice System

Ludwig and Mullainathan, 2021

Context

The criminal justice system relies heavily on human judgment in high-stakes decisions such as bail, sentencing, and parole. These decisions are often inconsistent, error-prone, and discriminatory.

There has been growing application of algorithms in a wide range of areas within the justice system. However, in practice, algorithms have often been found to be less helpful and in some cases more harmful.

In this paper the authors explore the inconsistency, error and discrimination of algorithms in the criminal justice system. They argue that the root cause of the failure of these algorithms is due to human decisions about how to build and deploy them.

Key Insights

Why algorithms fail

Human decisions
Humans make decisions about which outcome the algorithm should predict, and which variables predict that outcome.

Misaligned objectives
The objective of an algorithm is to optimise its prediction, but judges have other objectives: fairness. remorse, justice etc.

Inaccurate training data
Algorithms are trained using historical data that inherit past human bias and often previous decisions were made using information that is not captured in the data.

Presentation of results
The presentation of algorithmic results to the end user can influence their overall decision.

Missing governance
Regulation and laws are still catching up to AI, currently there is no evaluation of algorithms before deployment and no monitoring of bias.

The path forward

Incorporate real objectives & the right data
Recognise the objectives of decision-makers and the data they care about and incorporate these into algorithms.

Evaluate algorithm performance
Use econometric methods to account for human bias in data and missing data on the outcome that was not observed to evaluate the algorithm's predictions.

Education for end users
Help judges learn their comparative advantage and when to override algorithms.

Implications

- What are the objectives of algorithms you use and do these completely align with your organisation's objectives?
- How are you addressing missing or biased data?
- How do you currently test whether your algorithm improves outcomes pre and post deployment?
- What comparative advantage do people have over AI in your organisation? When is it right for them to override?

Ludwig, J., Mullainathan, S. (2021) 'Fragile Algorithms and Fallible Decision-Makers: Lessons from the Justice System'. *Journal of Economic Perspectives, 35*(4), 71-96. https://doi.org/10.1257/jep.35.4.71.