# Counterfactual fairness - The Alan Turing Institute

Algorithms are increasingly assisting in life-changing decisions, such as in parole hearings, loan applications, and university admissions. However, if the data used to train an algorithm contains societal biases against certain races, genders, or other demographic groups, then the algorithm will too. Using causal methods, researchers in this project are aiming to ensure algorithmic fairness by taking into account different social biases and compensating for them effectively.

Jump to: **Project aims** | **Applications** | **Recent updates** | **Disciplines & Techniques**

Joshua Loftus, Assistant Professor of Information at NYU Stern: "Fairness and causal inference are kind of dual problems. In causal inference, you try and figure out the effect of a certain variable, when it might be confounded with [i.e. both be affected by and affect] other things. In fairness, you're trying to make a certain variable not have any effect, when it too might potentially be confounded with other things." It's possible to utilise this idea to help better understand and mitigate issues with algorithmic decision-making systems.

Imagine a car insurance company wants to price insurance for car owners by predicting their accident rate with an algorithm. The company assumes that aggressive driving is linked both to drivers being more likely to have accidents and a preference for red cars.

Whilst it would seem that the company could use red car preference as a good way of predicting who will cause accidents, there could be other variables at play: for example, what if individuals of a particular race are more likely to drive red cars, but are no more likely to drive aggressively or have accidents? If, in an attempt to be fair, the company removed or ignored race from the decision-making process, they might only have directly observable factors, like red car preference, from which to make a decision, thereby including, rather removing, hidden racial biases.

What needs to happen instead is that when an algorithm is being designed to predict an outcome or make a decision, the variables used to make the prediction or decision need to be carefully assessed. Any

differences in these variables identified as being caused by sensitive factors, like race, then need to be cancelled out.

**Back to top**

Ensuring fairness using causal methods to produce 'counterfactually fair' algorithms, based on the idea a decision is fair towards an individual if the outcome is the same in reality as it would be in a 'counterfactual' world, in which the individual belongs to a different demographic.

The project has involved producing a technical set of guidelines for practitioners to use when creating or refining decision-making algorithms. These guidelines give practitioners an idea of how they should structure their problem, what is implied by their assumptions, and how they could evaluate it all using a causal model. The framework also calls for the need for decision-making algorithms to be designed with the input of expert knowledge about the situations the algorithms are being used in.

The research team are currently working and engaging with policy-makers, lawyers, and investigative journalists, whilst refining and improving their methods.

**Back to top**

One example of the team's work in action is with ProPublica, an American non-profit investigative newsroom, which has collected a big dataset from the use of an algorithm called COMPAS. The algorithm is used by judges and parole officers in the US for scoring criminal defendants' likelihood of reoffending. The team have worked with ProPublica to run their methods on the dataset in order to confirm systematic, racial bias is prevalent in the algorithm.

**Back to top**

The team presented their work in December 2017 at the Neural Information Processing Systems (NIPS) conference in California, one of the world's premier machine learning conferences.

In a year of record-breaking submissions for the conference, the team were invited to give an oral presentation of their research; one of only 40 of the 3,000 plus submitted papers selected for this kind of presentation.

Simon DeDeo, Assistant Professor in Social and Decision Sciences at Carnegie Mellon University and previously a Visiting Researcher at the Turing, was present at the conference and said: "A crucial part of how we make moral judgements is by talking about causes. The counterfactual fairness work is the first to show us how we might use

this to uncover injustices in algorithms".

**Back to top**

Causality

Algorithms

Machine learning

Ethics

**Back to top**