RUTH FREMSON/NYT/REDUX/EYEVINE

A migrant farm worker has her fingerprints scanned so that she can register for a national identity card in India.

The long road to fairer algorithms

Matt J. Kusner & Joshua R. Loftus

Build models that identify and mitigate the causes of discrimination.

An algorithm deployed across the United States is now known to underestimate the health needs of black patients[1]. The algorithm uses health-care costs as a proxy for health needs. But black patients' health-care costs have historically been lower because systemic racism has impeded their access to treatment – not because they are healthier.

This example illustrates how machine learning and artificial intelligence can maintain and amplify inequity. Most algorithms exploit crude correlations in data. Yet these correlations are often by-products of more salient social relationships (in the health-care example, treatment that is inaccessible is, by definition, cheaper), or chance occurrences that will not replicate.

To identify and mitigate discriminatory relationships in data, we need models that capture or account for the causal pathways that give rise to them. Here we outline what is required to build models that would allow us to explore ethical issues underlying seemingly objective analyses. Only by unearthing the true causes of discrimination can we build algorithms that correct for these.

Causal models

Models that account for causal pathways have three advantages. These 'causal models' are: tailored to the data at hand; allow us to account for quantities that aren't observed; and address shortcomings in current concepts of fairness (see 'Fairness four ways').

A causal model[2] represents how data are generated, and how variables might change in response to interventions. This can be shown as a graph in which each variable is a node and arrows represent the causal connections between them. Take, for example, a data set about who gets a visa to work in a country. There is information about the country each person comes from, the work they do, their religion and whether or not they obtained a visa (see 'Three causal tests', part 1).

This model says that the country of origin directly influences a person's religion and whether they obtain a visa; so, too, do religion and type of work. Having a causal model allows us to address questions related to ethics, such as does religion influence the visa process?

But because many different causal models could have led to a particular observed data set, it is not generally possible to identify the right causal model from that data set alone[3]. For example, without any extra assumptions, data generated from the causal graph described here could seem identical to those from a graph in which religion is no longer linked to visa granting. A modeller must therefore also leverage experiments and expert knowledge, and probe assumptions.

Experiments can help in identifying factors that affect fairness. For example, a modeller wishing to explore whether ethnicity would affect treatment recommendations made online by health-care professionals could create two patient profiles that differ only in some respect that relates to ethnicity. For instance, one profile could have a name common to Americans of Chinese descent, and the other a name common to Americans of African descent. If the treatment recommendations are the same, then names can be ruled out as a source of bias, and the model can be stress-tested in another way.

Few aspects of a deep, multifaceted concept can be tested as easily as changing a name. This means that experimental evidence can underestimate the effects of discrimination. Integration of expert knowledge, particularly from the social sciences and including qualitative methods, can help to overcome such limitations. This knowledge can be used to, for example, inform the modeller of variables that might be influential but unobserved (lighter circles in 'Three causal tests'), or to determine where to put arrows.

Assumptions about unobserved variables that might alter the predictions of a model need to be clearly stated. This is particularly important when experiments cannot be run or more detailed expert knowledge is not available. For example, if 'health-care access' is not observed in a model attempting to predict 'health need', then it is crucial to identify any potential impacts it might have on 'health costs' as well as how it is affected by 'ethnicity'.

This need for context and metadata makes causal models harder to build than non-causal ones. It can also make them a more powerful way to explore ethical questions.

Three tests

Causal models can test the fairness of predictive algorithms in three ways.

Counterfactuals. A causal model allows us to ask and answer questions such as 'Had the past been different, would the present or future have changed?' In the visa example (see 'Three causal tests', part 1), algorithmic biases could be smoked out by tweaking parts of the model to explore, for instance: 'Had individual X been Christian, would this algorithm have granted them a visa?' A researcher could then identify what pieces of information an algorithm could use to achieve counterfactual fairness[4]: the algorithm's output would not change regardless of the individual's religion. For example, if the algorithm used just work and not country of origin or religion, it would satisfy counterfactual fairness.

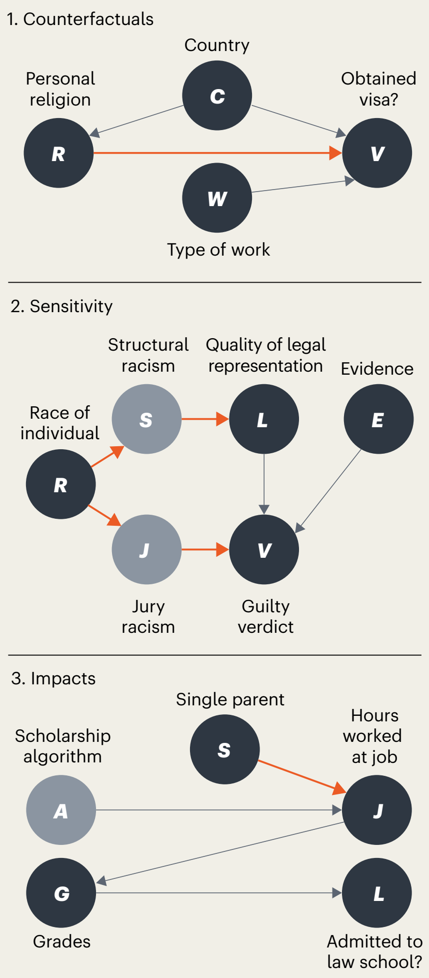
Sensitivity. In many settings, unknowns alter knowns — data we can observe are influenced by data we cannot. Consider a causal model for a trial setting (see 'Three causal tests', part 2).

This model shows how two independent sets of unobserved quantities, structural racism and jury racism, can unfairly lead to a guilty verdict. Although researchers often cannot precisely identify unobserved variables, they can reason about how sensitive a model is to them. For instance, they can explore how sensitive our estimate of the causal link between legal representation and guilty verdict is to different levels of jury racism. Simulations of the worst-case bias scenarios (that is, when jury racism is highest) can then be used to alter jury selection to minimize the bias.

Impacts. Data-driven decisions can have long-term consequences and spillover effects. These effects might not be obvious, especially in the standard machine-learning paradigm of predicting one short-term outcome. But carefully designed causal models can help researchers to use 'interventions' to probe the ripple effects of decisions far into the future[5,6]. For instance, the models can help regulatory agencies to understand how changing a scholarship algorithm influences who is accepted into law school (see 'Three causal tests', part 3). In this example, a single parent might need a scholarship so that they can reduce the hours they need to spend at a job, leaving them more time for study. That boosts their grades and therefore influences their chances of being admitted to law school. This complex chain can be explored using causal models.

THREE CAUSAL TESTS

Algorithmic fairness can be examined in different ways.


1. Counterfactuals

2. Sensitivity

3. Impacts

Five steps

Causal models are powerful tools, but they must be used appropriately. They are only models, and will thus fail to capture important aspects of the real world. Here we offer some guidelines on using them wisely.

Collaborate across fields. Researchers in statistics and machine learning need to know more about the causes of unfairness in society. They should work closely with those in disciplines such as law, social sciences and the humanities. This will help them to incorporate the context of the data used to train the algorithms. For example, scholars should meet at interdisciplinary workshops and conferences. One such is next year's Association for Computing and Machinery (ACM) conference on Fairness, Accountability and Transparency to derive a set of causal models for setting bail price and for immigration decisions.

A great example of such collaborations is one between information scientist Solon Barocas at Cornell University in Ithaca, New York, and attorney Andrew Selbst at the Data & Society Research Institute in New York City. They described how current law is unable to deal with algorithmic bias[7]. Partly in response to this work, machine-learning researchers have launched a large subfield, known as algorithmic fairness, that looks into ways of removing bias from data. And we and other researchers now use causal models to quantify discrimination due to data.

Partner with stakeholders. Predictive algorithms should be developed with people they are likely to affect. Stakeholders are best placed to provide input on difficult ethical questions and historical context. One example is the work by statistician Kristian Lum at the Human Rights Data Analysis Group in San Francisco, California, which investigates criminal-justice algorithms[8]. Such algorithms decide whether to detain or release arrested individuals and how high to set their bail, yet they are known to be biased. Lum has invited people affected by such decisions to speak at academic conferences attended by people who research these algorithms. This has led to closer collaboration, including the tutorial 'Understanding the context and consequences of pre-trial detention' presented at the 2018 ACM conference on Fairness, Accountability and Transparency in New York. So far, most stakeholder work has focused on criminal justice. Another setting that would benefit from it is mortgage lending.

We propose that a rotating, interdisciplinary panel of stakeholders investigates the impacts of algorithmic decisions, for example as part of a new international regulatory institute.

Make the workforce equitable. Women and people from minority groups are under-represented in the fields of statistics and machine learning. This directly contributes to the creation of unfair algorithms. For example, if facial detection software struggles to detect faces of black people[9], it is likely the algorithm was trained largely on data representing white people. Initiatives such as Black in AI (go.nature.com/38pbcaa) or Women in Machine Learning

Fairness four ways

A flurry of work has conceptualized fairness. Here are some of the most popular, and ways in which causal models offer alternatives.

Fairness through unawareness[12]. This method works by removing any data that are considered *prima facie* to be unfair. For example, for an algorithm used by judges making parole decisions, fairness through unawareness could dictate that data on ethnic origin should be removed when training this algorithm, whereas data on the number of previous offences can be used. But most data are biased. For instance, number of previous offences can bear the stamp of historical racial bias in policing, as can the use of plea bargaining (pleading guilty being more likely to reduce a sentence than arguing innocence)[13]. This can leave researchers with a hard choice: either remove all data or keep biased data.

Alternatively, causal models can directly quantify how data are biased.

Demographic parity[14]. A predictive algorithm satisfies demographic parity if, on average, it gives the same predictions to different groups. For example, a university-admissions algorithm would satisfy demographic parity for gender if 50% of its offers went to women and 50% to men. It is currently more common in law to relax demographic parity so that predictions aren't necessarily equal, but are not too imbalanced. Specifically, the US Equal Employment Opportunity Commission states that fair employment should satisfy the 80% rule: the acceptance rate for any group should be no less than 80% of that of the highest-accepted group. For instance, if 25% of women were offered jobs, and this is the highest acceptance rate, then at least 20% of men must be offered jobs[4]. One criticism of demographic parity is that it might not make sense to use it in certain settings, such as a fair arrest rate for violent crimes (men are significantly more likely to commit acts of violence)[15].

Instead, one could require that counterfactual versions of the same individual should get the same prediction[4].

Equality of opportunity[16]. This is the principle of giving the same beneficial predictions to individuals in each group. Consider a predictive algorithm that grants loans only to individuals who have paid back previous loans. It satisfies 'disability-based equality of opportunity' if it grants loans to the same percentage of individuals who both pay back and have a disability as it does to those who pay back and who do not have a disability. However, being able to pay back a loan in the first place can be affected by bias: discriminatory employers might be less likely to hire a person with a disability, which can make it harder for that person to pay back a loan. This societal unfairness is not captured by equality of opportunity.

A causal model could be used to quantify the bias and estimate an unbiased version of loan repayment.

Individual fairness[17]. This concept states that similar individuals should get similar predictions. If two people are alike except for their sexual orientation, say, an algorithm that displays job advertisements should display the same jobs to both. The main issue with this concept is how to define similar. In this example, training data will probably have been distorted by the fact that one in five individuals from sexual or gender minorities report discrimination against them in hiring, promotions and pay[18]. Thus similarity is hard to define, which makes individual fairness hard to use in practice.

In causal modelling, counterfactuals offer a natural way to define a similar individual. **M.J.K. & J.R.L.**

(go.nature.com/2s5km5g) are positive steps.

And we can go further. Causal models can themselves help to address the field's 'pipeline problem' by identifying where unfairness enters the process and which interventions can increase the participation of under-represented groups without shifting the burden to extra work for role models in those groups. Academic institutions should critically evaluate and use these models for fairer admissions in fields related to artificial intelligence.

Identify when algorithms are inappropriate. Statistics and machine learning are not all-powerful. Some problems should not be solved by expanding data-gathering capabilities and automating decisions. For example, a more accurate model for predictive policing won't solve many of the ethical concerns related to the criminal legal system. In fact, these methods can mask structural issues, including the fact that many neighbourhoods are policed by people who do not live in them[10]. This disconnect means that police officers might not be invested in the community they police or the people they arrest.

There are red flags when demographics, such as ethnic origin, influence nearly every piece of information in a causal graph, or when previous attempts to address a bias failed because people strategically changed behaviours in response. In these cases, an algorithmic solution would paper over a system that needs fundamental change.

Foment criticism. A vibrant culture of feedback is essential. Researchers need to continually question their models, evaluation techniques and assumptions. Useful as causal models are, they should be scrutinized intensely: bad models can make discrimination worse[11]. At the very least, a scientist should check whether a model has the right data to make causal claims, and how much these claims would change when the assumptions are relaxed.

Algorithms are increasingly used to make potentially life-changing decisions about people. By using causal models to formalize our understanding of discrimination, we must build these algorithms to respect the ethical standards required of human decision makers.

The authors

Matt J. Kusner is an associate professor in the Department of Computer Science at University College London, and a fellow at the Alan Turing Institute, London, UK. **Joshua R. Loftus** is an assistant professor in the Department of Technology, Operations, and Statistics at New York University, New York, USA.
e-mails: matt.kusner@gmail.com; loftus@nyu.edu

1. Obermeyer, Z. *et al. Science* **366**, 447–453 (2019).
2. Pearl, J. *Causality: Models, Reasoning, and Inference* (Cambridge Univ. Press, 2000).
3. Spirtes, P. *et al. Causation, Prediction, and Search* (MIT Press, 2000).
4. Kusner, M. J., Loftus, J., Russell, C. & Silva, R. In *Advances in Neural Information Processing Systems* 4066–4076 (MIT Press, 2017).
5. Liu, L. T. *et al.* In *International Conference on Machine Learning* 3150–3158 (ACM, 2018).
6. Kusner, M., Russell, C., Loftus, J. & Silva, R. *Proc. Machine Learning Res.* **97**, 3591–3600 (2019).
7. Barocas, S. & Selbst, A. D. *Calif. L. Rev.* **104**, 671 (2016).
8. Lum, K. *Nature Hum. Behav.* **1**, 0141 (2017).
9. Simon, M. 'HP looking into claim webcams can't see black people.' (CNN Tech, 23 December 2009).
10. McManus, H. D. *et al. Race Justice* https://doi.org/10.1177/2153368719849486 (2019).
11. Kilbertus, N. *et al.* 'The Sensitivity of Counterfactual Fairness to Unmeasured Confounding'. In *Uncertainty in Artificial Intelligence* (AUAI, 2019).
12. Grgic-Hlaca, N. *et al.* 'The case for process fairness in learning: Feature selection for fair decision making.' *NeurIPS Symposium on Machine Learning and the Law* (2016).
13. Wilford, M. M. & Khairalla, A. in *Social Sciences Contributions to the Real Legal System* Ch. 7, 132 (2019).
14. Zafar, M. B., Valera, I., Rogriguez, M. G. & Gummadi, K. P. In *Artificial Intelligence and Statistics* 962–970 (2017).
15. Dobash, R. E., Dobash, R. P., Cavanagh, K. & Lewis, R. *Violence Against Women* **10**, 577–605 (2004).
16. Hardt, M., Price, E. & Srebro, N. 'Equality of opportunity in supervised learning'. In *Advances in Neural Information Processing Systems* 3315–3323 (2016).
17. Dwork, C. *et al.* 'Fairness through awareness'. In *Proc. 3rd Innov. Theoret. Comp. Sci. Conf.* 214–226 (2012).
18. Pizer, J. C. *et al. Loy. LAL Rev.* **45**, 715 (2011).

Corrections

The long road to fairer algorithms

This comment omitted an affiliation for Matt J. Kusner. He is also a fellow at the Alan Turing Institute in London. It also gave the wrong author name for reference 2. The author is J. Pearl, not P. Judea.