Detecting Bias: Does an Algorithm Have to Be Transparent in Order to Be Fair?

William Seymour

University of Oxford, Oxford, UK
{william.seymour@cs.ox.ac.uk}

Abstract. The most commonly cited solution to problems surrounding algorithmic fairness is increased transparency. But how do we reconcile this point of view with the state of the art? Many of the most effective modern machine learning methods (such as neural networks) can have millions of variables, defying human understanding. This paper decomposes the quest for transparency and examines two of the options available using technical examples. By considering some of the current uses of machine learning and using human decision making as a null hypothesis, I suggest that pursuing transparent *outcomes* is the way forward, with the quest for transparent algorithms being a lost cause.

Introduction

Recent investigations into the fairness of algorithms have intensified the call for machine learning methods that are transparent. Unless an algorithm is transparent, so the argument goes, then how are we to know if it is fair? But this approach comes with a problem: many machine learning methods are useful precisely *because* they work in a way which is alien to conscious human reasoning. Thus, we place ourselves in the position of having to choose between a more limited (and potentially less effective) set of algorithms that work in ways that we can understand, and those which are better suited to the task at hand but cannot easily be explained. To clarify, this paper is concerned with the use of transparency as a tool for auditing and communicating decisions, rather than debate over the higher level 'transparency ideal', or harmful/obstructive uses of transparency as described by [1, 2].

This paper will discuss the arguments for and against transparency as a design requirement of machine learning algorithms. Firstly, I will break down what we mean when we talk about fairness and transparency, before considering arguments and examples from both sides of the discussion. I will cover two different black box techniques that provide interpretable explanations about algorithmic decisions—local explanations and statistical analyis—as well as some of the problems associated with each of these techniques. The techniques listed are by no means exhaustive and are meant to represent different styles that can be used to generate explanations. To conclude, there will be a discussion on the role that transparency might play in the future of machine learning.

What Do We Mean by Transparency?

Since transparency in this context is rooted in fairness, perhaps a better starting point would be to ask what we mean by fairness. A dauntingly complex question in itself, most people would consider approaches that 'treat similar people in similar ways' to be fair. These often coalesce along lines of protected characteristics (such as race and gender), as these are where the most glaring problems are often to be found. These characteristics are often expected to be excluded from the decision making process even if they are statistically related to its outcome.

But problems arise when a philosophical definition of fairness is translated into a set of statistical rules against which an algorithm is to be compared. There are multiple perpendicular axes against which one can judge an algorithm, and the best fit will vary based on the context in which the algorithm is used. Examples include predictive parity, error rate balance, and statistical parity to name a few [3]. To further muddy the waters, it is possible to draw a distinction between *process* fairness (the actual process of making a decision) and *outcome* fairness (the perceived fairness of a decision itself) [4]. It is possible for an algorithm with low process fairness (e.g. including race as a factor in decision making) to exhibit high output fairness (e.g. ensuring similar levels of false positives across racial groups).

As for the term transparency, I refer to information available about an algorithm that details part of its decision making process or information about the decisions it makes, which can be interpreted by a human being. Depending on the context, this could be a data scientist, policy maker, or even a member of the public. Interpretability is a key requirement here, ensuring that published data do actually aid our understanding of algorithmic processes.

As we are concerned about investigating fairness, it makes sense to think of two types of transparency corresponding to those for fairness: process transparency (how much we understand about the internal state of an algorithm) and outcome transparency (how much we understand about the decisions, and patterns in decisions, made by an algorithm). This distinction is important, as while there exist tools that can achieve some level of outcome transparency for all algorithms, only certain types of algorithm exhibit process transparency.

Method I: Local Explanations

The first method we consider is a black box method of explaining individual decisions. Local explanations work by sampling decisions from the problem domain weighted by proximity to the instance being explained. These samples are then used to construct a new model that accurately reflects the local decision boundary of the algorithm. For non-trivial algorithms, the local model will be a bad fit for other inputs, as global decision boundaries will be of a higher dimension than the local one (see Figure 1).

An example of this would be online content moderation. If a user has submitted a post which is deemed by an algorithm to be too toxic, we might want to explain

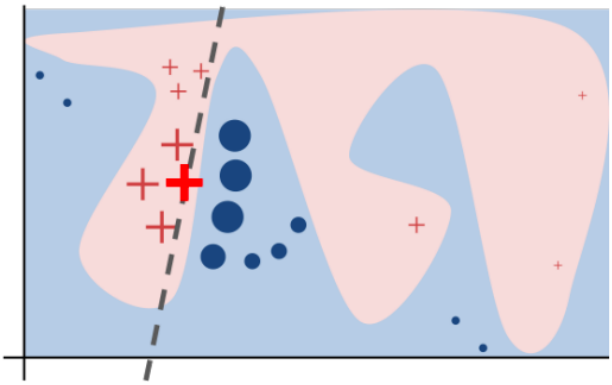

Fig. 1. Depiction of a local decision boundary (dotted line) of the bold cross derived from sampled decisions (shapes) weighted by their distance from the decision being explained. The global decision boundary is represented by the pink and blue background. Reproduced from [5] with permission.

to them which parts of their message caused the algorithm to reject it. For the input sentence

> "idiots. backward thinking people. nationalists. not accepting facts. susceptible to lies"[1]

a local explanation might reveal that the words "idiots", and "nationalists" are the greatest factors contributing to the message being flagged as toxic. This is not to say that all messages containing the word "nationalists" are toxic, but that the word is considered problematic in this context.

Here we have produced an interpretable explanation without knowing anything about how the algorithm operates—we can say that local explanations provide evidence for outcome fairness. By looking at these explanations for decisions a system makes, we have enough information to conclude that a decision was unfair because it violates our definition of fairness as described above. This is a good start to our goal of auditing for fairness.

Moving From Local to Global

Local explanations do a good job of informing users of the main factors behind the decisions they are subject to, but they fall short of providing assurance that the system as a whole operates fairly. In order for this to happen, one needs to be able to create a mental model of the system which is functionally close enough to the original that one can predict what it will do (or at least believe that

[1] Taken from the list of examples on the Google Perspective API home page at `https://www.perspectiveapi.com/`

its reasoning will be of sufficient quality). Because local explanations consider only facets of the *current* decision, they do not reveal much about the wider reasoning that pervades an algorithm. While of great use to an individual who is concerned about a decision concerning themselves, they are much less useful to an auditor who is seeking assurance that the algorithm as a whole is fair. A handful of randomly chosen samples being satisfactory does not give sufficient assurance that all answers will satisfy a set of fairness criteria. This highlights the distinction drawn earlier between local and global fairness guarantees.

Perhaps then, explanations for audits need to operate at a higher level than local explanations. But then we encounter the problem that the high dimensionality of non-trivial models means that global explanations must be simplified to the point of absurdity in order to be intelligible. If explanations can be thought of as "a three way trade off between the quality of the approximation vs. the ease of understanding the function and the size of the domain for which the approximation is valid" [6], then do we risk going so far towards the scale end of the spectrum that we must abandon our hopes of arriving at an answer which is also understandable and accurate?

Method II: Statistical Analysis

Given these problems it is perhaps questionable as to whether any scheme which only considers individual decisions can ever be sufficient to determine if an algorithm is fair or not. When considering higher level explanations of algorithms we find that statistical analysis can offer us the reassurance (or otherwise) that we desire about an algorithm, taking into accounts trends across entire groups of users rather than being limited to individual circumstances.

Statistical analysis is another black box method, and often takes the form of calculating information about particular groups of users and how they are dealt with by the algorithm. By comparing accuracies and error rates between groups it is possible to identify systemic mistreatment. Explaining these findings is often simple, given most people's intuitive understanding of accuracy and false positives/negatives (see Figure 2).

Lies, Damned Lies, and Statistics

One trap that exists when performing statistical analysis is that due to the aforementioned multitude of ways one can express statistical fairness it is almost always possible to present evidence of compliance *and* noncompliance. This is because many types of statistical fairness are inherently incompatible with each other: altering the classifier to increase fairness along one axis will always decrease it in another.

In the wake of Machine Bias [7], ProPublica and Northepoint argued that the COMPAS algorithm was unfair and fair, respectively. Both parties were technically correct. These explanations are thus only valid when paired with background knowledge in data science and ethics, and may not suitable for

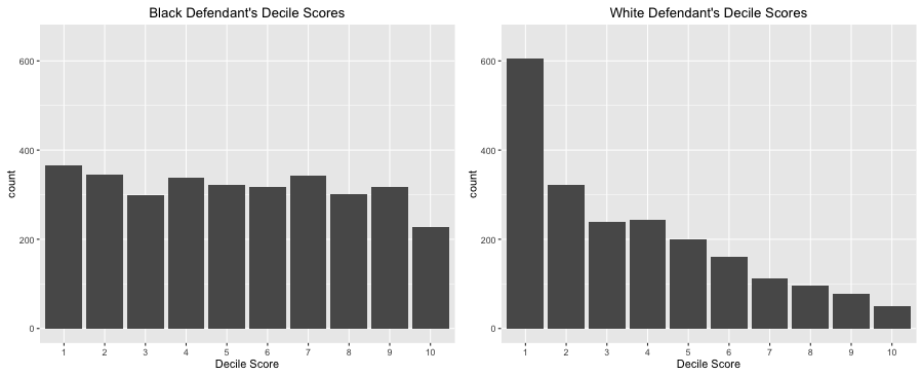

Fig. 2. A comparison of decile risk scores between black and white prisoners assessed by the a recidivism algorithm. A score of 1 represents the lowest risk, and 10 the highest risk. One can clearly see that white prisoners are much more likely to be classified as low risk. Generated from code published by [7].

presentation to the general public—doing so could lead to a reduction in trust of machine learning techniques, especially if the presented facts are used to support previously held beliefs which are incorrect [2].

Another issue is that all of the methods that provide interpretable decisions inevitably present reasoning that correlates with a decision making algorithm but is not causally related to its output. In these cases if the algorithms internals are indeed intractable then it will remain impossible to ever prove a causal link between the explanation system and the algorithm itself. This is not an insurmountable problem, by its nature all machine learning deals with correlations, but it needs to be understood that using black box analysis techniques is not enough to *guarantee* that a system is fair unless the entire problem domain is exhaustively searched. For any model big enough to require auditing this will be impossible.

Discussion

The point that becomes clear as we look at the realities surrounding transparency in machine learning is that exclusively pursuing understandable and/or open source algorithms is infeasible. When reviewing even a moderately-sized code base, it quickly becomes apparent that issues of transparency and interpretability cannot be resolved simply by making computer code available [8]. With a caveat for certain contexts, we need to be able to deal with algorithms that are not inherently transparent.

Put another way, industry players are incentivised to use the machine learning techniques that are best for profits, a decision which almost always favours efficacy over interpretability. Given this, we need to consider techniques that can

be applied to methods where the raw form of the model defies our understanding, such as neural networks.

The position I advocate for here is not that we should give up completely on pursuing transparency, but that we need to be clearer about what we are seeking. By failing to differentiate between process and outcome transparency we run the risk of intractable algorithms being used as an excuse for opaque and potentially unfair decision making.

At the same time, it is important to understand the epistemological implications that come from using correlation-based methods to provide transparency. However, this is already something that is being dealt with when it comes to algorithmic decisions themselves. If the rest of the community can tackle the misguided use of algorithmic 'evidence', then it is surely also possible to do the same with transparency.

Ultimately it is up to us to decide in each case whether the correlation-focused evidence we can generate about an algorithm is sufficient to draw conclusions about its fairness or unfairness. It is helpful to frame the question in the context of the alternative, which is human-led decision making. It is no secret that decisions made by people can occasionally be opaque and prone to bias [8], and using this human baseline as a null hypothesis reminds us that the goal of our quest for transparency should be for machines exceed our own capabilities, not to obtain perfection.

A realistic approach would be to use both types of technique (white and black box) in tandem, analysing the inner workings of simpler components where possible and utilising second hand explanations and analysis otherwise. We should remember that transparency can appear as a panacea for ethical issues arising from new technologies, and that the case of machine learning is unlikely to be any different [9]. That it is difficult to analyse the inner workings of particular techniques will not slow or prevent their uptake, and it is increasingly clear that there is a public and regulatory appetite for more accountable machine learning systems. Therefore, going forward we need to be focussed on the attainable if we are to effectively hold algorithm developers and their algorithms to account.

References

1. Ananny, M., Crawford, K.: Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. new media & society (2016)
2. Flyverbom, M.: Transparency: Mediation and the management of visibilities. International Journal of Communication **10**(1) (2016) 110–122
3. Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data **5**(2) (2017) 153–163
4. Grgic-Hlaca, N., Zafar, M.B., Gummadi, K.P., Weller, A.: The case for process fairness in learning: Feature selection for fair decision making. In: NIPS Symposium on Machine Learning and the Law. (2016)
5. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should I trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM (2016) 1135–1144

6. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the GDPR. (2017)
7. Larson, J., Mattu, S., Kirchner, L., Angwin, J.: How we analyzed the COMPAS recidivism algorithm. ProPublica (5 2016) (2016)
8. The Royal Society: Machine learning. Technical report, The Royal Society (2017)
9. Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S., Floridi, L.: The ethics of algorithms: Mapping the debate. Big Data & Society **3**(2) (2016)