Lernende Systeme

GERMANY'S PLATFORM FOR ARTIFICIAL INTELLIGENCE

Artificial Intelligence and Discrimination

White Paper by Susanne Beck et al.
Working Group IT Security, Privacy, Legal and Ethical Framework

Executive Summary

Artificial Intelligence (AI) is already being used far more widely today than we might initially expect, and the associated potential for discrimination is not always obvious. Although people themselves are guilty of unjustifiable discrimination at times, they perceive the decisions taken by computer programs and software solutions to often be factual, objective and neutral. However, in reality, AI-based systems sometimes make decisions that are problematic, discriminatory or that draw distinctions without good reason.[1] Many software systems explicitly or implicitly comprise a set of social rules for controlling behaviour, whether in the form of regulations, transactions and coordination, or access and usage rights. First and foremost, they are an effective technical means of putting systems of rules into practice. Consequently, self-learning systems have the potential to not just adopt pre-existing discrimination, but even to enhance it.

For example, in the United States, algorithms are used to identify the likelihood that the defendant in a criminal trial will reoffend.[2] These algorithms use different types of data to calculate a value that is supposed to give the judge a guide to the likelihood the defendant will go on to commit another offence. However, the algorithm is trained primarily on the basis of historical data (e.g. criminal statistics), which are based on statistical rather than causal correlations. The result is that defendants from sections of the population that have tended to appear in court more frequently than others in the past (such as ethnic minorities or groups with below-average financial means) receive worse

1 An initial overview of potentially overlooked negative consequences in the usage of Artificial Intelligence can be found in "Atlas of Automation" by AlgorithmWatch (AlgorithmWatch 2019).
2 See Angwin et al. 2016 for an overview of this practice.

prognoses. Since the judge's ruling is based partly on this calculation, defendants are immediately disadvantaged if they belong to one of these groups. Consequently, using these algorithms reinforces pre-existing distortions.

The quandary regarding the potential for discrimination when using Artificial Intelligence is part of a wider debate about the development and application of AI and its limits. Accordingly, this issue is also addressed in the German Federal Government's AI strategy and by its Data Ethics Commission and the German Parliament's Study Commission on Artificial Intelligence. At a European level, the European Commission's High-Level Expert Group on Artificial Intelligence has indicated in its "Ethics Guidelines for Trustworthy Artificial Intelligence" that self-learning systems ought to be free from discrimination. A number of companies have also already acknowledged this and have made appropriate voluntary commitments or have set up special ethics boards.

The Law and Ethics subgroup of the working group IT Security, Privacy, Legal and Ethical Framework of Plattform Lernende Systeme would like to contribute to this ongoing debate with this paper. The authors aim to first set out the different aspects of discrimination before then examining technological solutions[3] and focussing on social aspects. This will highlight the aspects of discrimination that must be addressed as part of a wider social dialogue and the institutions that can be helpful in doing this. The focus is on systems that make or suggest decisions that primarily affect people, their access to services and goods or their opportunities to participate in society. The paper shows that not all differentiation is unjustified, but that discrimination exists where there is no justification for equal or unequal treatment. Input and training data are the primary sources of discrimination by self-learning systems, but the application's output also has a role to play. The biggest challenges when seeking to eradicate discrimination from AI applications lie in a lack of transparency in algorithms, their continuous process of learning, the lack of neutrality in data and unclear responsibilities.

The authors single out the following lines of approach for the development of non-discriminatory self-learning systems:

Explainability and scrutiny

Decisions taken by AI should be traceable. If they are, it may be possible to accept certain forms of discrimination by AI. However, this raises more problems besides the technical challenges that are set out. System transparency is not an end in itself, considering commercial secrets are also important for technological advances. It is important to clarify just how transparent technology needs to be and to whom – and it may be appropriate for an independent institution to make those decisions.

An independent body could be set up to act as a representative for citizens who could potentially suffer discrimination and who, with a socially disadvantaged background, would usually find it difficult to exercise their rights. This body would be tasked with monitoring and evaluating the outputs of self-learning systems. It would use clearly defined instruments and principles to check the plausibility of results and the explanations issued by the systems

3 The report from the Dagstuhl Seminar 16291 (Abiteboul et al. 2016) and the Data Responsibly project can be consulted for an initial overview of technological solutions for ethical AI applications.

themselves. This raises the issue of the considerable speed at which systems change or new systems are brought into use. However, that is precisely why it is important to consider steps to ensure the (training) data and the methods used can be scrutinised externally and that test cases can be continuously improved. All the same, all this should take place with the understanding that Machine Learning can only observe and record correlations and not causal relationships. Machine Learning then takes places on the basis of these correlations. This fact makes it more difficult for a third, neutral body to monitor and check developments independently.

Ongoing staff training is also essential in businesses and public organisations that use these systems. Furthermore, besides setting up such a supervisory body, it is important to require the manufacturer or operator of such systems to undertake continuous follow-up monitoring, as certain types of discrimination only become evident during use. Should any discrimination come to light further down the line, the manufacturer or operator would be held responsible for rectifying the situation or liable for the damages caused by the discrimination.

Selection of criteria

To avoid discrimination, any traits that society regards as discriminatory in the given context (such as ethnicity, for example), ought to be removed from the input for Machine-Learning processes. Admittedly, this does not resolve the problem that many traits can act as proxies for another (Harcourt 2010). For example, the software used to assess the probability of a defendant reoffending did not include "race" among its input variables, but still effectively discriminates on the basis of this variable. That is because other proxy variables (such as place of residence, financial situation, etc.) can be used to draw conclusions about the original discriminating variable. The likely success of efforts to clean data before using it as input is disputed (Kilbertus et al. 2017, Doshi-Velez/Kim 2017) and, similarly, it is worthwhile discussing how to evaluate blatantly discriminatory output that it is clearly not based on discriminatory input.

In general, this approach requires a consensus on which criteria are discriminatory and which currently unforeseeable correlations we view as acceptable or unacceptable. The difficulty in achieving this consensus is evident in the protracted debate surrounding Article 3 of the Basic Law for the Federal Republic of Germany. This Article covers equality before the law, gender equality and the stipulation that no person shall be given an advantage or disadvantage because of various traits such as gender, parentage, faith or disability. The increased usage of self-learning systems will reinvigorate these debates and make it all the more important to find a consensus. It is also conceivable that – instead of trying to assess all categorisations in advance – institutions could be created that would be legitimately entitled to make these material decisions.

Making fair treatment the objective of Machine Learning

Another option would be to make fair treatment itself the objective of Machine-Learning processes. This would shift the focus away from achieving the most efficient or accurate classifications possible, and toward enabling the fairest possible classifications. This approach to fairness is being explored as part of research into Artificial Intelligence. However, these experiments have uncovered a quandary that is of fundamental importance to informational

research. Our ideas of what is "just" or "fair" cannot be formalised in terms of their complexity, which means it is not easy to make them an aim of Machine Learning. An attempt by Kleinberg et al. to do just that highlighted three ways that fairness could be formally expressed (2016):

- The forecast ought to be **"well calibrated"**. If an algorithm predicts that a certain characteristic will apply to a group with a specific probability, such as 0.1, then a proportion of the group that corresponds with this probability should demonstrate this characteristic – in this case one tenth.

- If there are several groups among the individuals who have been classified, such as men and women, then one could demand that there is a **"balance for the positive class"**. The average probability for a characteristic that is allocated to people who actually do possess that characteristic should be the same in every group. This ensures that no one group is classified with above-average false-positives.

- Similarly, it is possible to demand a **"balance for the negative class"**. The average probability for a characteristic that is allocated to people who do not possess that characteristic should be the same in every group. This ensures that no one group is classified with above-average false-negatives.

Kleinberg et al. are showing that it is not possible to satisfy all three of these intuitively correct characterisations perfectly and simultaneously. This clarifies the inherent technical limits of the approach toward integrating fairness into Machine Learning.

Effective legal protection and law enforcement

In addition to the aforementioned approaches to developing solutions, it is also important that affected parties are put in a position to defend their rights. This does not mean that they have to be trained as experts in Artificial Intelligence, but rather that affected parties should be notified of their rights and given the opportunity to actually exercise them. This includes the option of defending those rights in court. The financial cost for this should be kept as low as possible. The option of taking out insurance against discrimination by self-learning systems is potentially another option that should be made possible. State authorities have the task of countering illegal discrimination by self-learning systems and yet, in all these cases, regulation should be kept within sensible limits and over-regulation should be avoided.

Imprint
Editor: Lernende Systeme – Germany's Platform for Artificial Intelligence | Managing Office | c/o acatech | Karolinenplatz 4 | D-80333 München | kontakt@plattform-lernende-systeme.de | www.plattform-lernende-systeme.de | Follow us on Twitter: @LernendeSysteme | Status: June 2019 | Image credit: r.classen / Shutterstock

This executive summary is based on the white paper Artificial Intelligence and discrimination – challenges and approaches to developing solutions, Munich, 2019. The authors are members of the working group IT Security, Privacy, Legal and Ethical Framework of Plattform Lernende Systeme. The original version of this publication is available at: https://www.plattform-lernende-systeme.de/publikationen.html

SPONSORED BY THE

Federal Ministry of Education and Research

acatech
NATIONAL ACADEMY OF SCIENCE AND ENGINEERING