Center for
Big Data Statistics

Working paper

Fair algorithms in context

Working paper no.: 05-20

Rik Helwegen
Barteld Braaksma

May 2020

Abstract

Machine Learning has become more powerful over the past decade, sparking an expansion of new applications. Some of these applications fall within the social domain, in which models based on data profiles can have a significant impact on the life of individuals. In order to prevent unwanted discrimination in these models, different methods have been proposed within the field of algorithmic fairness. The present paper aims to provide context for fairness methods, connecting technical research with public debate and practical considerations. The goal of algorithmic fairness is further defined, and separated from related problems which might lead to confusion in ongoing debate. A fairness method which relies on causality theory is discussed, making our recent technical research available for a non-specialised audience. Finally, fair algorithms are considered from the perspective of practical deployment, locating challenges in bringing the theory into practice.

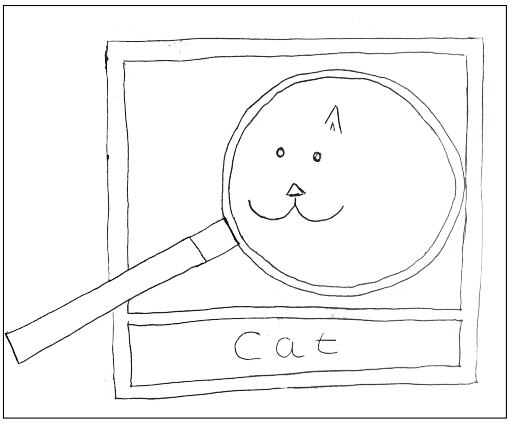
Contents

1 Introduction

Decisions in our society are increasingly supported by **Artificial Intelligence** (AI). In this work we refer to AI as computers doing 'intelligent' tasks. This can for example be determining good moves in a game of chess, or recognising cats in images. Over the past decade, the performance of AI systems has improved significantly (LeCun et al., 2015). Main drivers behind this development are increasing amounts of data, increasingly powerful computers, and better techniques to train AI systems. That last aspect includes machine learning, which is the subfield of AI we will focus on.

In **machine learning**, the approach to making a computer more intelligent is to have it learn by itself. This can be preferable over writing out all instructions by hand, because it is more time-efficient, and allows to find patterns in complex information. Software which has obtained intelligence through machine learning is what we call a machine learning model. The developer of such a model has to provide very clear instructions on what the model should learn, and which learning procedure it should follow. Besides instructions on how to learn, a model needs to be fed by information. Information in digital form, which is what we call *data*. For the example of teaching a computer to recognise cats in images, we need images with their corresponding label, indicating if an image contains a cat or not. The machine learning model looks for patterns in the data, starting off with a random focus over the pixels. As a learning procedure, the model needs to improve the focus patterns in an iterative process, alternating between prediction and evaluation based update steps. Correct predictions indicate that the model is looking at relevant patterns, so focus on those patterns should be increased. For wrong predictions, the model should put less attention on the used patterns.

1.1 Machine Learning uses patterns in the data to form predictions, for example, it can be trained to recognise features of cats in images



The model is trained by repeating this principle for many pictures. A possible outcome pattern could be pixel values which indicate pointy ears in combination with eyes and a nose. The model uses this pattern to *differentiate* pictures with cats from other pictures. If the instructions are not provided clear enough, the model will not come up with such a pattern. The idea that AI can have free thought, or freedom in what it is going to optimise for, is far from today's practical reality. However, it does happen that machine learning methods find patterns for a particular problem which we find hard to understand as humans, or which we just never thought of before. Silver et al., 2017 introduced a machine learning model for the game of

chess, which surprised the chess community with moves that where uncommon in professional chess but then turned out to be very strategic, outsmarting the world champions.

The current expansion of machine learning **applications** can be understood from the increasing amount of data which is stored in our society. This data can be used to inform decision makers with condensed historical knowledge. According to the promising prospects of machine learning, this can be done very efficient in terms of labour, and can yield outcomes which exceed the quality of alternative options. One specific kind of applications is making predictions based on the data profiles of individuals. A data profile consists of attributes, such as a person's age, gender and activity history. In order to make meaningful predictions, models need to find differences between data profiles. This way, people can be grouped together such that information about the group can be used to make predictions for an individual. Many products which use personalised predictions are already embraced by society. Think of the Google search engine, Spotify music recommendations, or the Facebook news feed. The term *algorithm* in this work, refers to such applications of machine learning models.

The fact that machine learning models act on differences between people makes the notion of **Fairness** relevant. In some domains, like music recommendations, the effects of such differentiation has relatively low impact on people's lives. On the other hand, high impact decisions are also increasingly being supported by algorithms. For example, companies use machine learning to decide which people would be good applicants for job vacancies, banks can predict credit scores in order to decide whether someone should be given a loan, and law enforcers can perform their work more efficiently by using risk models. When the predictions of such models are being influenced by *sensitive attributes*, like gender or ethnicity, this can constitute unwanted discrimination from a legal or ethical perspective. Sensitive attributes are defined as variables which are deemed unfair to base differentiation on.

The question 'What is fair?' is central to the debate on algorithmic fairness. A question in which the technical, legal and ethical perspectives are bound to come together. This working paper is published alongside a technical paper[1)] in which a fairness method is proposed, building on state-of-the-art techniques. Such papers require a certain level of specialisation to be read and understood. The overarching goal of this working paper is to:

1. Connect technical literature and public debate on algorithmic fairness
2. Open up the findings of our research to a broad, non-specialised audience
3. Discuss practical considerations for using fairness methods

The content of this work should be read as one perspective on a multidisciplinary problem. Remarks, additions or points to discuss are most welcome. The work continues in section 2 by further isolating and explaining the problem which fair algorithms aim to solve. In section 3, a high-level overview of the FairTrade method is provided, a method which aims to prevent unwanted discrimination in machine learning models. Section 4 discusses practical considerations for working with fairness methods, and section 5 summarises and concludes this working paper.

1) The work 'Causality-based Fairness using Variational Inference' by Rik Helwegen, Christos Louizos and Patrick Forré is currently under review at a machine learning conference, and will be shared on arXiv.org after the end of the reviewing period in June 2020.

2 Problem description

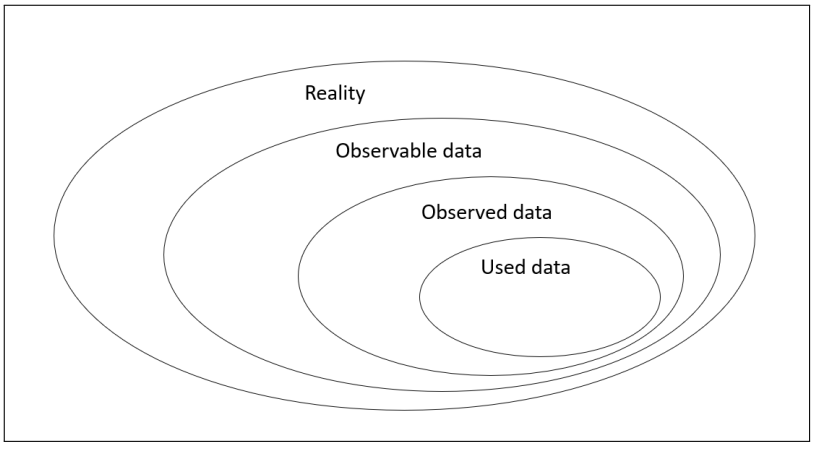
This section describes the problem of unwanted discrimination, and explains how it can be further defined. This step is needed in order to have constructive debate on algorithmic fairness, to avoid that the debate, as often, ends in confusion about language and the achievable goal. In short, algorithmic fairness refers to the goal of excluding a given selection of sensitive information from prediction models in a fair way.

2.1 Inequality in the data

One cause that an algorithm shows unwanted discrimination is that it uses unfair information to base its predictions on. The unfair information is often defined in the form of sensitive attributes, for which gender and ethnicity are examples. The attributes separate groups, such as males and females. If the data for males and females are further indistinguishable (the conditional distributions are the same), a well optimised machine learning model will in general give equal treatment to the groups. Likewise, if two sensitive groups obtain unequal treatment by a model, this is often due to inequality in the data. Other possible reasons for unwanted discrimination can include mistakenly interpreting the data or model outcomes, a wrong conceptualisation of reality, inequality caused by the structure or incorrect optimisation of the model. In this subsection, we focus on understanding inequality in the data, independently of other causes for unequal outcomes.

In order to understand the *origin* of inequality in the data, it is useful to visualise the different steps going from reality to a set of used data. Figure 1 is a schematic view on this process, which is discussed using an example on weather forecast data below.

2.1 Levels of data. The reality might contain sensitive inequalities, and each step towards a usable dataset can introduce new or stronger inequalities.



On the highest level, there is the data describing reality, the existing weather in this example. Sometimes it rains and it is cold, while at other times it is warm and sunny. Many characteristics of the weather can be conceptualised and measured, such as humidity and temperature. This creates a subset of *observable* data. It would be impossible however to measure everywhere, all the time, as we only have limited capacity of measurement instruments. We might be more

inclined to keep track of the temperature in large cities compared to areas where nobody lives. Hence, the second layer of *observed* data becomes only a subset of the observable data. Out of the set of observed data, again a subset is selected in order to train prediction models. The *used* data might focus on a particular area, or be restricted to recent observations.

Moving back to the social domain, each step of going to a subsequent layer of data can introduce sensitive inequalities. This can be illustrated by a police department assigning more attention to specific neighbourhoods in a city. Having a limited number of police officers, the department might focus attention towards neighbourhoods in which more criminality is expected. One result of this is that the *observed* criminality does not properly represent the criminality in *reality* throughout the city. If a sensitive group is overrepresented in one of those focus neighbourhoods, the **inequality in used data** may incorrectly describe their share in criminality. Next to inequality created by (unconscious) bias, there might have been **inequality in reality** as well, a second type of inequality. For example, social factors might cause groups of one gender to be more involved in criminality compared to a group of a different gender. This means that, even if the used training data perfectly represents reality, the sensitive group will obtain a different algorithmic treatment.

In general, people agree that models should be corrected for a biased sample. This point of view is not only in favour of making fair predictions, but also helps in making predictions more accurate. The second type of inequality, existing in reality, is less obvious to correct for. Doing so would create a less 'objective' representation, because it no longer represents reality. Nevertheless, in different cases it is argued to carry out such corrections, for specific issues prioritising equality over approximating reality. For example, it can be legally bounded to equalise insurance premiums for different sensitive groups, even though there might be evidence suggesting to do otherwise. Enforcing equal treatment can be a way to steer towards equality in reality over time. Giving people the same opportunities might change the equality in reality in the long run. If such corrections indeed force a model to level out differences which are present in reality, the predictions will become more equal but less accurate. Whether this is a fair thing to do is a political or ethical choice, and is seldom addressed in technical fairness literature. A related discussion which is generally lacking in machine learning fairness literature is how to handle cases in which both types of inequality are likely to be present. Many of the context cases in the public debate are likely to be of this kind. Admittedly, it is difficult and sometimes impossible to distinguish between these two types of inequalities. If practically feasible, taking a proper random sample to analyse bias in the data can be an important tool to analyse different inequalities.

The difference between these two kinds of inequality causes difficulty in the debate on algorithmic fairness. Much critique on algorithms focuses on biased collection of data, causing the machine learning models to make unfair predictions. In response, people working in law enforcement, recruiting or credit scoring can argue that there is undeniable inequality between sensitive groups, which is reflected by the algorithm. Both of these standpoints can be true simultaneously, which needs to be understood well before groups on different sides of the discussion can come together.

2.2 Fairness criteria

The field of algorithmic fairness has quickly become more popular over the past few years, showing a steep increase in technical publications analysing the problem of unwanted discrimination. In order to operationalise fairness in the domain of machine learning, it requires a

definition which can be measured in terms of data or model predictions. Hence, a main development in algorithmic fairness takes place around statistically defining fairness criteria, also called fairness metrics in the literature.

Probably the simplest criterion is named *Fairness by Unawareness*. The criterion states that any model which does not use sensitive attributes is called fair. It has the important shortcoming that it does not take into account possible *proxy variables* of the sensitive attribute. Proxy variables contain information about a different variable because they are related. For example, a variable on Dutch language capabilities can be a proxy for a variable on migration background in the Netherlands. If only migration background is a sensitive attribute, and the unawareness principle is applied, the model might still include migration information through the variable on language capabilities. Another proposed criterion is *Demographic Parity*, according to which a model is fair if the predictive distribution is the same for different sensitive groups. Slightly modifying this idea, *Equalized Odds* proposes that the rate of correct predictions is the same for different sensitive groups. Both of these criteria say something about the equality between entire groups. This leaves open the possibility for unwanted discrimination on an individual level, which might level out when comparing statistics at group level. In reply to this problem, the fairness criterion of *Individual Fairness* was proposed. This criterion states that a model is fair if the predictions are similar for people that are similar apart from sensitive information. The prerequisite of a similarity measure is the major problem with this criterion, as it will be hard to design but has direct influence on the measured fairness.

An important next step in defining fairness was taken by Kusner et al. (2017) when introducing *Counterfactual Fairness*. The authors note that many of the problems in the aforementioned fairness metrics are due to a limited understanding of relations between variables in the data profiles. In order to solve this, they combined the fields of **Causality theory** and algorithmic fairness. Causality theory is a topic with a long history, and is approached differently per discipline of research. For machine learning, a major source of definitions for causality theory is the work by Pearl (2009). One central concept introduced by Pearl is the *Intervention*. We speak of an intervention in a system if a variable is forced to take on a specific value, which might influence the value of other variables. This is an interesting concept because it separates correlation based statistics and causal relations. Consider measurements of altitude and temperature for cities around the world. Comparing their values would probably show a relation between the two variables, in which cities on higher locations have a lower average temperature, in accordance with our expectations from theory. A possible intervention would be to heat up an entire city, raising the temperature. If we would measure the altitude of this city before and after the intervention, there will be no apparent change. Another (hypothetical) intervention would be to pick up a city at the beach, and put it on top of a mountain. After doing so, there will be a measurable difference in the average temperature for this city, caused by the difference in altitude. What this experiment shows is that there is a causal relationship from height to temperature, but not the other way around. This is something which cannot be described by statistical correlation only, as correlation is a symmetric measure, not accounting for directed relations between variables. The idea behind counterfactual fairness is that an intervention on a sensitive attribute should not influence the model outcomes for an observed person, also not when this intervention causes other variables in the data profile to change. This creates the intuitive criteria that a sensitive variable is not allowed to be the cause for discrimination in machine learning models.

Ultimately, these criteria aim to statistically formalise expressions of fairness, which have their origin in a social or philosophical discipline. This creates a **language problem**. The original

description of fairness might be hard to understand for a computer scientist, and the fairness criteria might be hard to understand for experts in social studies or philosophy. The intuitive notion of causality-based fairness criteria help to create a common language between disciplines.

2.3 Related but different problems

In this section, a number of problems are discussed which are closely related to the algorithmic fairness problem, but of a different nature. To start with, the prerequisite of being provided with a selection of sensitive information raises the question who should decide on what information is fair to use. Important to consider here is that the predictive power of a model can decrease when excluding (sensitive) information. Marking information as 'sensitive' might therefore conflict with the goal of the user or developer to obtain the highest possible accuracy. This can motivate to put this decision at the level of a democratically chosen person or entity, a councillor of a city for example. Another challenge is to embed an algorithm in the process and context for which the model has been built, how will the model cooperate with the existing workforce? This can affect the autonomy of the employees concerned, it might change the number of jobs, change the kind of people needed, and raise new regulatory considerations. In addition, infringement of privacy is an issue which can directly conflict with fairness methods. As most fairness methods rely on (the absence of) sensitive relations in the outcomes, usage of attributes like ethnicity are likely to be required for implementation. While this is needed in the perspective of discrimination prevention, it might be harmful from the perspective of privacy protection.

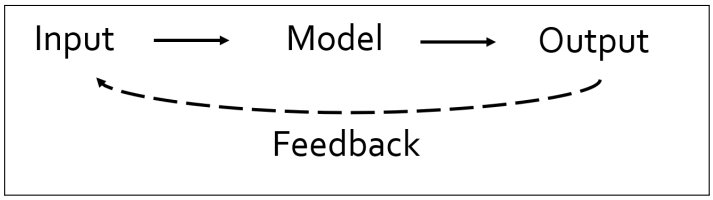
The above issues indicate that the usage of algorithms implies interaction with a larger socio-economic system, creating effects on a level which is not necessarily overseen by the developer of the machine learning model. Next to these higher-level problems, there are more technical problems as well to take into account which are relevant but distinct from algorithmic fairness. First, explainability of machine learning models is a separate problem often desired in the social domain. This refers to the understanding of how and why a model has come to its prediction for a specific individual. If a model is completely explainable, it is often relatively straightforward to exclude unfair information. On the other hand, a model which adheres to fairness criteria does not necessarily need to be explainable.

Another closely related technical problem is inconsistency in the model quality for different sensitive groups. This may cause unfair results, but falls outside the scope of the considered problem of algorithmic fairness. For example, there has been debate on computer vision algorithms which did not perform well on people with a darker skin tone. An important explanation for this was an insufficient number of people of colour in the training data. As a result, the model was unable to learn to properly recognise particular groups of people, leading to undesired outcomes. Most fairness methods focus on discrimination as a result of information stored in the data. In contrast, this example shows unfairness as a result of a shortage of data.

Finally, feedback loops are another important issue when considering the effects of algorithms on society. As the prediction of an algorithm might influence the newly obtained observations, retraining of models can obtain self-amplifying effects. Consider the example of a bank which offers loans to people who obtain a high algorithmic credit score. The group which is rated as financially unstable is denied access to financial products, which might deteriorate their financial situation even more. If the model is retrained after some period of time, the people denied

in the first model will be given an even lower credit score this time, closing the loop. Feedback loops are not limited to algorithms, and naturally occur in all sorts of situations. However, the fact that algorithms can be scaled up easily and may affect large groups in an uncontrolled way makes the problem specifically relevant in this context.

2.2 Feedback loops are constituted when the output of a model has influence on what comes into the model, this can create a self-amplifying effect.



These related problems, which are non-exhaustive, are not within scope of the fairness methods considered here. That raises the question: is it worth while to study algorithmic fairness in itself, if so many problems are left out of scope? Given the increasing importance of information-based decision making, it is essential to progress in the field of algorithmic fairness. Doing that for many problems simultaneously is difficult and confusing, hence we attempt here to take a small step for one of the challenges surrounding the deployment of algorithms in society.

3 FairTrade method

Causality-based fairness criteria are intuitive to understand, and solve several of the problems which came up in earlier fairness criteria. Hence, it would be valuable to integrate such criteria in machine learning applications. Different studies in algorithmic fairness propose approaches to do so, among which the research by Kusner et al. (2017), Loftus et al. (2018), Nabi and Shpitser (2018) and Chiappa (2019). The present paper tries to provide some context for such methods, and is published alongside a technical paper introducing the FairTrade method. The FairTrade method aims to provide a practical approach to integrate the latest causality-based fairness criteria into real-life applications of machine learning algorithms. The three steps which constitutes the method are described here on an abstract level, and we refer to the technical paper for more details.

Step 1: Model causal structure
Causal modelling depends on assumptions on relations between the variables of interest. The first step of the FairTrade method is therefore concerned with making such assumptions. The assumed relations are based on domain knowledge, using existing theory on the topic of interest, and qualitative analysis of the problem at hand. For the example above, one can assume that temperature is influenced by the altitude of measurement.

Step 2: Learn causal mechanisms
In the second step of the FairTrade method, the relational assumptions are combined with data in order to estimate the form and strength of the relations. At this point, there is a trade-off to be considered on the complexity of the estimation method. Less complex methods, like a

regression model, have the benefit of being clearly interpretable and more easy to adjust for unwanted effects. On the other hand, only simple patterns can be recognised by such models, such as linear relations. This can be sufficient, especially when the data is well structured, but might also decrease predictive power compared to using more advanced methods. The FairTrade method is especially suitable for large data profiles which may contain complex patterns that are valuable for prediction. The model is created by taking the relational assumptions explicitly into account, building on recent advances by Louizos et al (2017). The causal relations are obtained by having the model learn to recreate the observed data. A key aspect in which the FairTrade method improves on existing methods is by taking into account unobserved confounders. This term refers to factors which have influenced multiple of the observed variables, but are not observed themselves directly. For data profiles of individual persons, social economic status is an example.

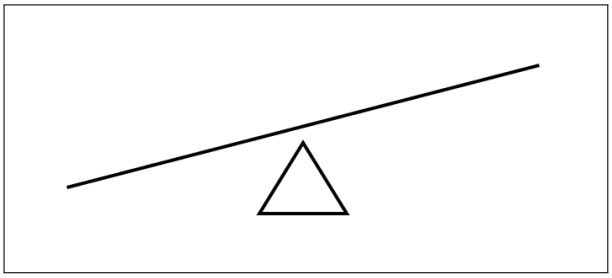
Step 3: Train fair prediction models

After obtaining the causal mechanisms, the resulting relational model is now a simplified representation of how the observed data has been created. The derived relations express how certain variables influenced others. At this point, we can train an auxiliary prediction model for which we exclude all the information (and proxy information) which is deemed unfair to use for prediction.

3.1 Trade-off between predictive power and equality

Two consequences of excluding information from a fairness perspective should be considered. First, for the FairTrade method in particular, the fairness constraints imply that the model is structured according to causal assumptions. This structuring itself is a restriction on the optimisation freedom for the model. The 'optimisation freedom' refers to the variety of patterns in the data the model can look at in order to predict its target. If this variety decreases, the model might lose performance because the optimal pattern is no longer within reach of the optimisation. The second consequence is that information is being deliberately excluded from the input of the model. Less information to base predictions on generally results in less accurate predictions.

3.1 When excluding information from prediction models to enforce equality between groups or individuals, this can come at the cost of prediction accuracy, hence establishing a trade-off between fairness and accuracy.



The cost of prediction accuracy when enforcing fairness establishes a trade-off. This *Fairness Trade-off* with accuracy is the origin of the FairTrade method name, and has been demonstrated in detailed experiments. To put this result in the perspective of the different layers of information (Figure 1), the trade-off can be shown within the subset of *observed data*. This is

done by keeping a part of the data apart during training, and then evaluating the trained model on this test set for different fairness constraints. If the method corrects inequalities which are actually present in society (*reality* level), the decrease in accuracy is expected to carry over in case the model will be deployed in practice. Interestingly, this does not hold for correcting inequalities which are a result of a biased sample, and hence exist in the *used data* but not in *reality*. For example, reconsider the police which does more focused screenings in one neighbourhood compared to others. This might cause inequality in the data, which is not present in reality for the city as a whole. If this inequality is corrected before applying a model to the entire city (*reality*), this correction can make the total accuracy increase, although the accuracy on the test set (*observed data*) decreased. The discussed trade-off puts emphasis on the question: *What are we optimising for?* Predictive power might be a too narrow objective for deployed algorithms with social consequences.

3.2 Limitations

The presented method has several limitations, from which the three most important ones are addressed here. First, the causal assumptions which are imposed on the model are very hard to verify, and contain the risk of bringing in new biases. This is an inherent problem for working with causality theory, and ongoing research aims to mitigate this (Forré and Mooij, 2019). A second limitation stems from the inference method used in the second step, obtaining the causal mechanisms. The model requires an approximation method in order for the optimisation to remain practically feasible. Third, a limitation is that we will in general not be able to evaluate the counterfactual estimates of the model. Changing the sensitive attribute for a person is something which can not be done in practice, and therefore the estimates cannot be compared against true observations. This implies the FairTrade method needs to be applied with care; human judgement remains necessary to interpret and validate the results.

3.3 Experiments

In order to test the FairTrade method in practice, a number of experiments have been conducted using realistic scenarios. One of these experiments focused on predicting risk models for unlawful social welfare, in which, as a research objective, ethnicity is marked as the sensitive variable. This experiment used a data set made available for this study (research purposes only) by Statistics Netherlands (CBS). In the closed CBS IT environment, over 90.000 detailed profiles were constructed of people who received social welfare or had been convicted for unlawfully receiving social welfare in the past. Domain knowledge was built up by relevant literature and by conducting interviews with experienced social researchers at CBS and the municipality of Amsterdam. The causal relations were estimated according to the FairTrade method, and prediction models were trained under different levels of fairness constraints. The trade-off between fairness and accuracy has been shown this way for a large scale real-life data experiment within a relevant context. The technical paper provides further explanation on the experiments and results.

4 Practical deployment of fair algorithms

As mentioned above, several methods are being developed to mitigate the risk of unwanted discrimination in machine learning models. The obvious goal of these methods is to be applied in practical scenarios to prevent harm being done to individuals. However, often there is a gap between the theoretical methods and the practical deployment of fair algorithms. This section aims to identify and discuss some of the problems which constitute this gap.

If an organisation wants to deploy fairness aware machine learning methods, a first prerequisite is to have a clean, sufficiently rich and well understood data set that can be used to obtain the desired causal relations. The data profiles need to be qualitatively detailed, available in high volume, and they need to include information about the sensitive attributes. Too little data increases the errors by the model, which can be especially harmful if it concerns insufficient data for one sensitive group. A second challenge is to bring together the right team of people for the job. As most fairness methods rely on algorithms combined with statistical criteria, technical people, trained in AI, statistical inference and IT are needed to build and maintain the model. Furthermore, sufficient domain knowledge is essential within the team. This domain knowledge is essential to understand how the data is observed, how variables might interact, and where possible sensitive inequality appears. The fact that fairness methods require to actually work with sensitive data can also cause extra demand for legal and communication skills in the team. Privacy regulations on sensitive variables are often strict, and their use requires proper interpetation and explanation towards users, impacted people and interested outsiders. Further down the process, it needs to be decided which level of fairness constraints the deployed model should aim at. This requires a fundamental understanding of the problem at hand, and oversight in the implied consequences of enforcing different constraints. One way to increase oversight in this is by building multiple models under various constraints, and comparing their outcomes. As mentioned before, the trade-off between equality and accuracy might be a reason to shift the ethical decision making away from the developers of the model itself. Finally, like with any new technology, embedding of the model in the working process can lead to a variety of challenges. For example, the work instructions might change for existing employees, with new technologies being integrated in daily activities. A level of understanding of the model and its fairness constraints are needed for successful cooperation in practice.

These challenges give some insight in what issues one has to face when trying to deploy theories on algorithmic fairness in practical cases. Further research on possible paths do deployment, ways to overcome challenges and shared experiences of fairness applications are of high value to increase accessibility of algorithmic fairness.

5 Conclusion

In summary, the increasing capabilities of AI creates reason to think about which ethical values we want to see reflected in our automated processes. A wide range of problems has to be considered. The present paper focuses on the prevention of unwanted discrimination from a

technical perspective; although we have tried to give a non-technical description. Difference in treatment can be unwanted because it is based on a sensitive attribute, such as ethnicity.

Inequalities in the data are a main source for unequal algorithmic treatment. Such inequality can stem from existing inequality in reality, but it can also be introduced during the process of obtaining a data set used for training. Different definitions of fairness have been proposed over the past several years, and one promising direction of research is the development of causality-based fairness criteria.

Causality increases the understanding of how unwanted effects might propagate through a data profile, making it possible to be more explicit which effects should be excluded from prediction models. In addition to this, the intuitive form of causality theory provides a perspective which allows discussion between disciplines. The FairTrade method proposes a way to incorporate causality-based fairness criteria into practical applications. Experiments show a trade-off between equality between groups and predictive performance. This trade-off puts new emphasis on a question with growing importance in the machine learning domain: What are we optimising for? A specific conclusion relates to the interaction between algorithms-based systems and human operations. As shown above, the current state of the art does not (yet) allow deployment of machine-learning algorithms that can run autonomously without human intervention.

References

Chiappa, Silvia (2019). "Path-specific counterfactual fairness". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Volume 33, pages 7801–7808.

Kusner, Matt J et al. (2017). "Counterfactual fairness". In: *Advances in Neural Information Processing Systems*, pages 4066–4076.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning". In: *nature* 521.7553, pages 436–444.

Loftus, Joshua R et al. (2018). "Causal reasoning for algorithmic fairness". In: *arXiv preprint arXiv:1805.05859*.

Louizos, Christos et al. (2017). "Causal effect inference with deep latent-variable models". In: *Advances in Neural Information Processing Systems*, pages 6446–6456.

Nabi, Razieh and Ilya Shpitser (2018). "Fair inference on outcomes". In: *Thirty-Second AAAI Conference on Artificial Intelligence*.

Pearl, Judea (2009). *Causality*. Cambridge university press.

Silver, David et al. (2017). "Mastering chess and shogi by self-play with a general reinforcement learning algorithm". In: *arXiv preprint arXiv:1712.01815*.

Acknowledgement

This working paper has been written in the context of a project on *fair algorithms*, in which Statistics Netherlands (CBS) collaborated with the municipality of Amsterdam, the University of Amsterdam, codefor.nl, the Association of Netherlands Municipalities (VNG) and other Dutch municipalities. The project is funded by the Ministry of the Interior and Kingdom Relations (BZK). We express our gratitude to Tamas Erkelens, Patrick Forré and Christos Louizos for their enthusiastic involvement and essential input throughout the project. Special gratitude to Gerhard Dekker, Matthijs Eggers, Sigrid van Hoek and Anna Mitriaieva as part of the great CBS team working on the fair algorithms project. Finally, the authors would like to thank Tobias van der Knaap and Leon Willenborg for reviewing earlier versions of the paper, and providing insightful comments and ideas.

Colophon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
Statistics Netherlands

Design
Edenspiekermann

Enquiries
Telephone: +31 88 570 70 70
Via contact form: www.cbs.nl/infoservice