The Accuracy Paradox of Algorithmic Classification

Fabian FISCHER

Institute for Visual Computing & Human-Centered Technology, Multidisciplinary Design & User Research, Vienna University of Technology, Austria

Abstract

In recent years, algorithmic classification based on machine learning techniques has been increasingly permeating our lives. With their increased ubiquity, negative social consequences have come to light. Among these consequences are ’unfair’ algorithms. This resulted in a large body of research tackling ’fairness’ of algorithms and related issues. Algorithms are frequently considered as unfair if they show diverging accuracies for different groups, with a particular focus on vulnerable groups, indicating a correlation between prediction and information about group membership.

In this paper I argue that, while this research contributes valuable insights, much of the research focuses a quantitative understanding of fairness which creates a very narrow focus. My argument builds on four pillars. First, much of the research on 'fairness' focuses on accuracy as basis for ’fairness’. Even though ’fairness’ can reduce the overall accuracy, this is seen as a limitation, implicitly aiming for high accuracy. Second, this focus is in line with other debates about algorithmic classification that focus on quantitative performance measures. Third, close attention on accuracy may be a pragmatic and well-intended stance for practitioners but can distract from problematizing the ’bigger picture’. Fourth, I argue that any classification produces a marginalized group, namely those that are misclassified. This marginalization increases with the classifier’s accuracy, and in tandem the ability of the affected to challenge the classification is diminished.

Combined, this leads to the situation that a focus on fairness and accuracy may weaken the position and agency of those being misclassified, paradoxically contradicting the promissory narrative of ’fixing’ algorithms through optimizing fairness and accuracy.

1 Introduction

As Machine Learning has started to increasingly permeate all aspects of our lives – particularly since the deep learning boom starting in 2012 – a sensitivity for potential downsides to machine learning has set hold in machine learning research and neighbouring fields, but also the wider public.

A particular set of issues has become popular in the past decade in debates of societal impacts of algorithmic systems, namely issues hinging on the terms ’fairness’, ’accountability’ and ’transparency’, frequently combined in the acronym ’FAT’. These topics have taken centre stage in a series of workshops and conferences such as Fairness, Accountability, and Transparency in Machine Learning (FAT ML) since 2014 and the ACM Conference on Fairness, Accountability, and Transparency (ACM FAT*) since 2018. A plethora of papers developing conceptualisations and mathematical formulations of fairness, as well as methods and tool kits to detect and remedy unfairness have been published. The research has a focus on, but is not limited to, machine learning.

Research and debates on these topics has also been noted in a wider public discourse, and by public bodies. This has, for example, resulted in the formation of the High-Level Expert Group on Artificial Intelligence, set up by the European Commission, which also tackles issues of ethics (High-Level Expert Group on Artificial Intelligence 2019). Furthermore, when new algorithmic systems are being introduced, particularly in public agencies and other areas with far-reaching consequences for affected people, these systems increasingly get discussed in terms of fairness, accountability and transparency.

I welcome this sensitivity for societal impacts of algorithmic systems. But in discussions of fairness, an emphasis is frequently put on a very narrow, quantitative understanding of fairness. I will show how this narrow definition of fairness is deeply entangled with notions of accuracy as a quantitative measure for algorithmic classification systems.

In the context of the ongoing digital transformation, algorithmic classification systems are becoming wide spread. In this context, accuracy is often used as a measure for the system’s quality, as exemplified by an OECD report on the use of statistical profiling – a special form of classification – in public employment services, where a lack of accuracy is highlighted as one potential shortcoming of these profiling systems (Desiere, Langenbucher, and Struyven 2019). More importantly, it is stated that "[p]ostimplementation, continuous evaluation and updates of the system based on feedback from all stakeholders will improve the system and its accuracy and will also help to build trust in it" (Desiere, Langenbucher, and Struyven 2019, 3).

This paper aims to contribute to the rich body of work within STS that studies classification and quantification. With Bowker and Star (1999) I argue that classification ist "not inherently a bad thing–indeed it is inescapable. But it *is* an ethical choice, and as such it is dangerous–not bad, but dangerous" (5–6). As I will discuss in detail in section 3, accuracy figures are frequently used as quantifications of quality, a process that abstracts and reduces (context) information to obtain representations (cf. Espeland and Stevens 1998; Porter 1995; Latour and Woolgar 1986).

In this conceptualizing piece point towards issues connected to the strong focus on accuracy, either directly or implicitly by focusing on a narrow understanding of fairness. First, I will introduce the concept of accuracy and some frequently used definitions of 'fairness' and their entanglement. Second I will investigate how the focus on accuracy may have come about. Third I will develop an argument why this focus on accuracy, and fairness in a narrow sense strongly limits the perspective of potential societal impact of algorithmic classification systems. Finally, I will argue that algorithmic classification produces its own marginalised individuals, namely those that get misclassified, and that this marginalisation intensifies with increased accuracy and may lead to a reduced agency of those who are affected. This, I argue, paradoxically contradicts the promissory narrative of improving, or 'fixing' algorithmic systems by optimizing fairness and accuracy.

2 Definitions of accuracy and fairness

Accuracy is a frequently used metric for algorithmic classification systems and often used to assess its quality. In this section, I will go into some detail how accuracy, and some related metrics, are intertwined with many (quantitative) approaches to 'fairness'.

A majority of machine learning systems are, on some level, classification systems. But not all classification systems are necessarily using machine learning. In computer science, an algorithmic classification system can also be called a classifier. Throughout this section I will illustrate the definition of accuracy by a cat picture classification example: An algorithm should classify a picture as either showing a cat, or as not showing a cat.

Accuracy in the context of algorithmic classification is a statistical measure that indicates the fraction of correct classifications. It is used to quantify the correctness of classifiers. The generalised calculation of accuracy is as follows

If the classification is binary, such as the cat picture example, the calculation can be reformulated as follows:

Where *True Positives* are in the above example cat pictures that correctly got classified as such. *True Negatives* are pictures that are not cat pictures and correctly got classified as non-cat pictures. *False Positives* are pictures that are classified as cat pictures but are actually not and *False Negatives* are cat pictures that haven't been classified correctly as cat picture.

In some fairness research, other measures play an important role, for example the **true positive rate** (Gajane and Pechenizkiy 2018; Wattenberg, Viégas, and Hardt 2016) that signifies how many have been correctly classified into one class: Sticking with the example of cat pictures, the true positive rate signifies how many of all cat pictures got correctly classified as cat pictures. This measure does not take into account how many non-cat pictures got incorrectly classified as cat pictures or correctly classified as non-cat pictures.

There are additional, closely related measures. While these differ in what they express exactly, they operate with similar concepts such as *True Positives* and *True Negatives*. More importantly, for my argument they are sufficiently close so that their function is interchangeable. All these measures aim to quantify the correctness of algorithmic classification, even though they emphasize different types of sources of misclassification. Common to all these measures is their quantitative nature and the aim to maximize these numbers: A 'good' algorithmic classifier shows high accuracy numbers, or a related metric. In the remainder of this work, I will use use the term accuracy as a shorthand for accuracy and these closely related correctness measures.

In the past decade, a multitude of definitions and formalisations of **fairness** have been used in the machine learning context. Frequently used are individual fairness, group fairness and equality of opportunity.

Individual fairness is when similar individuals get treated similarly (Friedler, Scheidegger, and Venkatasubramanian 2016). How 'similarity' can be defined is heavily dependent on the model used. In the context of college admission, for example, one approach could be to use the performance on standardized tests to measure similarity (Friedler, Scheidegger, and Venkatasubramanian 2016). This approach trusts that the model is adequate and ignores any structural effects that may lead to some groups performing better than others in that model. This formalization of fairness is closely related to accurate classifiers: Only accurate classifiers can treat similar individuals similarly.

Quite different to that is **group fairness**: Here, the population is divided into sub-populations based on some attributes (e.g. gendered or racialized attributes) and all groups are expected to be treated equally (Friedler, Scheidegger, and Venkatasubramanian 2016; Gajane and Pechenizkiy 2018). In the case of algorithmic classification, equal treatment is frequently understood as an algorithm that shows (almost) equal accuracy for each sub-population (corresponding to disparate impact (Friedler, Scheidegger, and Venkatasubramanian 2016)) or the true positive rate (corresponding to equal opportunity (Wattenberg, Viégas, and Hardt 2016; Gajane and Pechenizkiy 2018)). Here, again accuracy or a related metric is the foundation for checking if an algorithm is fair.

There are some additional formalisations with slightly different emphasis. Most operate on a group-level, and frequently groups identified through ’sensitive’ attributes, e.g., gendered and racialized attributes. Gajane & Pechenizkiy (2018) provide a concise overview over many formalizations. The aim of fairness research is, then, to prevent disadvantaging of groups, often explicitly of marginalised groups defined by ’sensitive’ attributes.

What is important to highlight is that narrow, quantitative formalizations of fairness are based on some correctness metric of the classifier that is either accuracy or a close relative. This is the first source of entanglement between accuracy and fairness in the narrow sense. Another way these two quantitative measures are entangled is that this approach to fairness also aims for high accuracy (see, e.g., (Kleinberg, Mullainathan, and Raghavan 2016)), sometimes noting that fairness may negatively impact the classifier’s accuracy, asking for a trade-off between accuracy and fairness (see, e.g., (Kearns et al. 2018)). As a consequence of this, algorithmic fairness discussed on quantitative terms is implicitly also discussing accuracy – reinforcing accuracy as a central aspect in the evaluation of algorithms, an aspect that I will focus on in the next section.

3 Accuracy as focal point

As I have described in the previous section, accuracy and a narrow definition of fairness are closely entangled. In this section I argue that this understanding of fairness is dominant in the fairness research, and consequently accuracy gets a strong emphasis in fairness research. In addition to that, I investigate other sources why accuracy and quantitative conceptions of fairness are in focus in many debates about algorithmic systems.

Addressed topic	Number of papers
Fairness	22
Accountability	2
Transparency	6

Number of papers addressing the topics from the FAT/ML 2018 Conference (n = 27)

Table 2: *Topics of FAT/ML 2018 papers*

First, even within the FAT discourse the focus on fairness is much stronger than on accountability and transparency. Table 2 shows the issues addressed in 27 papers presented at the FAT/ML 2018 Conference, showing a strong emphasis on the topic ’fairness’. Further examinations of the papers addressing fairness shows that few go beyond quantitative conceptualisations of fairness. Dobbe, Dean, Gilbert, & Kohli (2018) highlight issues of fairness that go beyond numeric qualities of fairness. Green (2018) provides a fundamental critique of the quantitative nature of machine learning. The increased use of machine learning, Green argues, grants "undue weight to quantified considerations at the expense of unquantified ones." (Green 2018, sec. 2.1, para. 1). Green continues to argue that some "aspects of society resist quantification" (Green 2018, sec. 2.1, para. 1).

Other papers I labelled as ’fairness’ research in Table 1 don’t directly use the term fairness but rather focus on bias or diversity and thus don’t neatly fit the above presented characterisation, either. For example, Ogunyale, Bryant, & Howard (2018) look at different perceptions of robots based on the different colorations of the robot.

Even though not all papers work with a narrow, quantitative conceptualisations of fairness, the majority of papers does, mostly by devising techniques and methods to ensure some that the algorithmic systems meets some accuracy-based notion of fairness. Other researches have a similar impression that a narrow, technical, quantitative conceptualisation of fairness prevails in fairness research (Selbst et al. 2019).

I want to identify some reasons why I think this quantitative, accuracy-based notion of fairness is at the focus of debate. They come from a range of directions.

One source for accuracy as focal point is an **optimization logic**. Data and computer scientists may simply be driven by a desire to improve everything (Morozov 2013). And there is hardly any easier way to quantify optimality than to increase a measure expressed in per-cents. This is true even though within computer science it is well known, that exclusively aiming for high accuracy numbers can have unwanted consequences. For example, optimizing for a given data set can mean that the machine learning algorithm is too narrowly learning characteristics of this particular data set and cannot generalise well to new data points. Picking up the cat picture example, if the data set used for learning always shows cats in a particular pose, it may actually learn the pose, not cats more generally. Still, this is seen as a risk to accurately classify new images without questioning the optimization logic itself.

Division of labor and competences within an organisation may also be a reason why data and computer scientists may focus on a rather narrow definition of fairness. Frequently, it is up to the management to formalize the problem at stake. Hence, defining, or questioning, the problem formulation is outside the data scientist’s sphere of influence. Similarly, the decision to solve a problem with algorithmic classification is often not up to them. In an attempt to make the best out of this situation, they may attempt to make the most fair system possible given the constraints by optimizing for fairness.

This ’not my department’ attitude is also a side-effect of a tendency for modularization and atomization of tasks when building algorithmic systems: Isolated units of work with minimal interaction with other parts of the system ensure on the one hand a divide-and-conquer approach to developing the system, but they also make it hard to assume a holistic view.

This also holds to the practice of out sourcing computer science work to external contractors, which can lead to a lack of understanding of the context by the contractor and a lack of understanding of the technical details on the side of the client.

An important source for the focus on fairness is legislation. While the exact definitions vary across countries and by context, compliance with various anti-discrimination laws is one motivator to investigate fairness (Gajane and Pechenizkiy 2018). **Liability concerns** are, then, a reason why businesses, management and public agencies are interested in fairness. Depending on the exact legal definitions, looking at accuracy figures (or closely related measures is a quantifiable way to demonstrate a fair classification system. This also explains why big corporations are paying so much attention to fairness: Being able to provide in this sense ’fair’ algorithmic solutions on the one hands can calm down any criticism and on the other hand makes it easier to enter new business fields where the legislation applies.

This leads to another potential reason why accuracy is getting so much attention: Drawing on a core STS concept, accuracy figures can take the function of inscription devices that provide "the focus of discussion about properties of the substance." (Latour and Woolgar 1986, 51), in this case about the purported quality of algorithmic classification. This is important, because the introduction of algorithmic systems can face resistance, both in the organizations where they are deployed and by those affected by the algorithmic classification. Accuracy as signifier for quality is in line with the practice to quantify qualities (Espeland and Stevens 1998; Porter 1995). Being able to refer to high accuracy is an easy way to **persuade critics** that the new system is doing a good job. For example, the OECD report on statistical profiling in public employment services states that the "usefulness and legitimacy of statistical profiling models hinge on model accuracy" (Desiere, Langenbucher, and Struyven 2019, 15).

At the same time, it is difficult to problematize the algorithmic system in other, more nuanced, more qualitative ways. It would be easiest to provide other kinds of figures to counter arguments based on accuracy or fairness numbers (Latour 1983). Which is then, maybe, one reason why controversies of algorithmic systems often focuses on this kind of figures, effectively establishing discourse coalitions (Hajer 2006) by sharing storylines about the problem at hand: Discrimination and fairness, expressed through figures.

Even if the accuracy is not perfectly high, it is easier to **promise accuracy improvements** in the future – a convenient way to take out some steam from the critique. This is, by the way, not necessarily simple rhetoric, but an honest promise. After all, research is making progress in machine learning – and this general progress can indeed lead to improvements in the local setting. These improvements are easily quantifiable, too: If the criticism is aimed at low numbers, presenting higher numbers can silence some of the critique. Contrary to improving this metric it would be much more challenging to attempt to change the system where the algorithmic system is embedded in in more qualitative, deeper ways.

Beyond one algorithmic system, accuracy as common measurement enables **quantifiable comparisons** across sites, and of alternate systems for one site (Desrosières 2010). It is easy to argue that an algorithm with high accuracy is better than one with lower accuracy – and ignore contextual contingencies. As Espeland and Stevens (1998) argue, the practice to transform qualities into quantities "is a way to reduce and simplify disparate information into numbers that can easily be compared" (316). An evaluative statement is much harder to make when comparing different organisational structures, processes, or types of data. By boasting high accuracy scores, one may be seen as a particularly splendid example of how a certain kind of system is done. Exemplified is this by the aforementioned OECD report, that compares several profiling systems using very different data sources and problem formulations, but evaluates them primarily via accuracy, lauding, for example, the Austrian system for its comparatively high accuracy.

4 Accuracy as distraction from the problem formulation

I have already touched how a focus on accuracy may function as distraction from other, maybe more important aspects of the algorithmic classification system. In this section I am going to extend these thoughts.

Focusing on accuracy can, at worst, not at all problematize the underlying data. As research on databases and, more recently, big data has plentifully showed, data are never 'raw', but always 'cooked' (Bowker 2005; boyd and Crawford 2012). Data are never 'just' 'out there' and ready to be discovered, but are always produced. And this usually with some goal in mind, i.e. what data is left out and what is produced and recorded is decided to support a particular goal.

Against this backdrop, a focus on accuracy and quantitative fairness may look at the data at hand and work with whatever is available. This may mean choosing a certain machine learning approach, this may mean taking only a subset of the available data points. The latter is particularly the case if analysis reveals that accuracy is different for certain protected attributes, or that some other variables correlate with protected attributes and using them can, too, lead to unfair classification.

Focusing on this particular perspective on fairness, one quickly looses a sense of the bigger picture. Namely, that any data set is not the same as the messy, contingent and contextual ’real life’ of the affected people. Data are always models, abstractions, or inscriptions. Consequently, any optimization with regards to fairness on this data does not ensure fairness in this messy ’real life’ – it is always an optimization on the model.

In a recent paper, Selbst, boyd, Friedler, Venkatasubramanian, & Vertesi (2019) have pointed out many issues related to this abstraction work done in computer science and assert that most papers dealing with fairness and machine learning "abstract away any context that surrounds this system." (Selbst et al. 2019, 59) The authors then continue to argue, similarly to my argument, that the technical systems are only a subsystem of more complex, contextual socio-technical systems, contributing many insights from Science & Technology Studies to the fairness debate. Importantly, they argue that abstraction is "taken as given and is rarely if ever interrogated for validity" (Selbst et al. 2019, 59)

In a similar fashion, even though not as elaborate, is Green’s argument that not all parts of society can be quantified and that the focus on quantification renders alternative approaches to solve societal issues invisible (Green 2018).

Barabas, Virza, Dinakar, Ito, & Zittrain (2018) tackle the application of risk assessment systems in the justice system. Risk assessment is a particular form of algorithmic classification, attempting to calculate and predict the risk a person may pose to society. The authors argue that debates about these systems’ fairness have hidden more fundamental issues with prediction in the justice system, arguing for the use of machine learning to identify causal "drivers of criminal behavior" (Barabas et al. 2018, 6).

In line with these researchers I argue that concentrating, researching, and debating fairness by means of accuracy and related measures distracts from problematizing the underlying model and problem formulation. I argue that the amount of attention given to research in this area may even be performative in a way to frame fairness as the main issue when discussing the societal impact of algorithmic classification systems. Or, as Barabas et al. have put it, this narrow focus prevents us from "ask[ing] harder questions" (2018, 7).

Continuing this thought further, creating ’fair’ algorithms to socially questionable models and problem formulations may actually stabilise and reinforce them as legitimate. Consequently, even critical data scientists who are slightly uneasy with the task they need to solve and hence set out to develop it in a fair way can eventually actively contribute to this legitimization.

5 Increasing accuracy intensifies marginalization and reduces agency of the misclassified

Until now I have shown that first, fairness research is largely done in a narrow quantitative sense and second, that implicitly or explicitly accuracy (or one of its close relatives) is seen as a quality measure for algorithmic classification and its optimization the goal. And, as I have stated, fairness aims at reducing the disadvantaging of usually marginalized groups. What is, however, not being discussed in fairness research, is that any algorithmic classification is producing its own marginalized group, namely those who get misclassified.

Fairness research tends to focus on well-established sub-populations defined by gendered or racialized attributes, by religion, or by socio-economic status, etc. Some of the fairness research addresses not disadvantaging these populations even if the sensitive attribute of individuals is not known. This can be due to legal issues of collecting these attributes in the first place or simply inability to get this information. Very little research also looks beyond these typical sub-populations and looks, e.g., at "socially meaningful subgroups" (Dwork and Ilvento 2018, sec. 2.2., para. 4) such as mothers. The majority focuses on protected attributes.

The marginalized group produced by misclassification is different to these notions of marginalizations. While the misclassification has to be based on some attributes, it is important to stress that this marginalized group is not necessarily related to the sub-populations defined via protected attributes. Implementing a fair algorithm as discussed above should actually prevent such correlations. The aforementioned work by Dwork & Ilvento (2018) is actually coming close to tackling this issue. Yet, the authors’ focus on

’socially meaningful’ groups assumes a pre-existing shared attribute and that these are considered ’socially meaningful’. Different to this, the group of misclassified individuals is produced *by the algorithmic classification system*. This system does, indeed decide on data, and hence attributes. But individuals from distinct ’socially meaningful’ groups may be misclassified on different grounds.

A fictitious example would be a facial recognition system that shows higher rates of misclassification if individuals wear some headwear (e.g. baseball caps), particular style of make-up, or simply have a particular posture which affects the angle of the face to the camera. These cases would hardly be known beforehand and don't relate to any 'socially meaningful' group. These higher rates of misclassification could lead to, e.g., more frequent checks by police, or more frequent denials of access (in case the facial recognition system is used as authentication).

The problem of misclassification is even more troubling if the system is producing predictions about the future, such as the system developed by the Austrian public employment service. This system attempts to predict the chances of job seekers to find a stable job in the future. Job seekers who are predicted to not find a stable job within the next two years will get access to different resources than those that are predicted to succeed. The corollary to this observation is that increasing the accuracy of an algorithmic classification system – which is, as I have outlined above, an overarching goal in machine learning research – is also intensifying the marginalization. The group of misclassified is literally increasingly marginal and the majority of correctly classified individuals is getting stronger. A consequence of this is that it will become harder for misclassified individuals to contest the algorithmic classification.

Let’s do a simple thought experiment: If it is known that the algorithmic system is wrong one third of the time, then it is easy to argue that the decision that affects an individual is incorrect. After all the odds are low. In contrast to this, it is harder to argue why the algorithmic system is wrong in a particular case. The likelihood for this to happen is small. This is why I draw a parallel to the popular claim of big data being superior to ’small’ data. This stance implicitly argues that "the volume of data adds to the weight of evidence" (Kitchin 2014, 135). I argue that, similarly, high accuracy of algorithmic classification adds to the weight to the individual classification.

Picking up the term ’evidence’ from Kitchin, there is another problem that may arise if an algorithmic classification system’s accuracy gets close to 100%. What if the algorithmic classification will, at some point, gain proof-like character? What if the algorithmic classification is seen as convincing evidence, if it gets recognized similarly to fingerprinting today (Cole 2009)?

The latter concern is not pulled out of thin air. In a recent media article about facial recognition to be used by the Austrian police, the head of the police records department states that ultimately courts will have to decide whether expert testimonies based on facial recognition will be accepted as evidence (Al-Youssef and Sulzbacher 2019). Even some fairness researchers explicitly point towards the potential usage of algorithmic classification "as evidence in legal proceedings" (Raff and Sylvester 2018, sec. 1, para. 1).

An important consequence of intensified marginalization is that the agency of misclassified individuals is reduced. First, as stated, the position to challenge the classification is weakened if the algorithmic classification system’s accuracy is high. The burden to show why the algorithm is wrong in a particular case is increasingly being shifted to the affected individual, not the operator of the algorithmic system. Second, simply due to fewer people affected by misclassification because the likelihood of getting in contact with other misclassified individuals is lower. This may hinder efforts to form self-help groups and organize from below.

6 Conclusion

In this work I have shown four issues. First, an overarching goal to increase accuracy guides the research of algorithmic classification systems. Second, the majority of research on fairness in machine learning is quantitatively oriented and tightly entangled with notions of accuracy or closely related measures. Third, as other critical researchers have been pointing out, focusing on this quantitative, narrow understanding of fairness renders important aspects beyond technical details invisible. This includes the problem formulation, issues of abstraction and modelling – both practices that reduce the complexity of social life – and the importance of social context. Finally, I have argued that the strong focus on accuracy ignores the issue of misclassification and that increasing accuracy paradoxically intensifies the marginalization of misclassified individuals.

Importantly, there's a frequent narrative in debates about algorithmic classification systems that promises that with increasing accuracy, many issues with algorithmic classifications, e.g. for decision making, will be solved, or at least reduced. A majority of fairness research takes up on this narrative, extending the issue so that algorithmic classification should work with equal or at least similar accuracy, or related performance metrics, across sub-populations. Neither approach focuses on problems that can arise for those that still get misclassified – which will occur even if the classification is close to perfect and 'fair' in quantitative terms.

Additionally, endeavours to improve an algorithmic classification system’s accuracy and fairness both signify quality and hence contribute to strengthening its legitimation. This stronger legitimation will, in turn, make it harder for individuals to object the classification.

While some work has started addressing marginalization brought about by the algorithmic classification itself, the vast majority of fairness research is still focusing on well-defined known sub-populations defined by protected attributes.

With my work I want to point to two open questions for future research. First, there’s a need to investigate marginalization practices that are done by algorithmic classification systems. This is in contrast to historical marginalization based on, e.g., racialized attributes. Second, the guiding theme in many the debates on machine learning that optimizing accuracy of algorithmic classification has to be problematized. How misclassification can be handled in a meaningful and ’fair’ way has to be addressed – this issue cannot be ’optimized away’, since all algorithmic classification must operate on a simplified abstraction of a complex and messy ’real world’. One proposition to address this is to focus more on meaningful channels to dissent the classification (cf. Skirpan and Gorelick 2017), and more generally on processes instead of results. Depending on how these processes look, they could provide agency to the affected individuals and could re-introduce context information that got lost in the necessary abstraction of the algorithmic system.

Acknowledgements

I would like to thank Doris Allhutter for extensive feedback to a very early draft of this paper, the participants of the STS Conference Graz for their feedback to my talk that this paper accompanies, and the anonymous reviewers for their valuable reviews.

References

Al-Youssef, Muzayen, and Markus Sulzbacher. 2019. “Polizei startet im Herbst Gesichtserkennung mit öffentlichen Kameras.” DER STANDARD. April 18, 2019. https://derstandard.at/2000101679966/Polizei-startet-im-Dezember-mit-Gesichtserkennung.

Barabas, Chelsea, Madars Virza, Karthik Dinakar, Joichi Ito, and Jonathan Zittrain. 2018. “Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment.” In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, edited by Sorelle A. Friedler and Christo Wilson, 81:62–76. Proceedings of Machine Learning Research. New York, NY, USA: PMLR. http://proceedings.mlr.press/v81/barabas18a.html.

Bowker, G. C. 2005. *Memory Practices in the Sciences*. Cambridge, Massachusetts: MIT Press.

Bowker, G. C. and Star, S. L. 1999. *Sorting things out. Classification and its consequences.* Cambridge, Massachusetts: MIT Press.

boyd, danah, and Kate Crawford. 2012. “CRITICAL QUESTIONS FOR BIG DATA.” *Information, Communication & Society* 15 (5): 662–79. https://doi.org/10.1080/1369118X.2012.678878.

Cole, Simon A. 2009. *Suspect Identities: A History of Fingerprinting and Criminal Identification*. Cambridge, Massachusetts: Harvard University Press.

Desiere, Sam, Kristine Langenbucher, and Ludo Struyven. 2019. “Statistical Profiling in Public Employment Services: An international comparison.” *OECD Social, Employment and Migration Working Papers*, no. 224 (February). Paris: OECD Publishing.

Desrosières, Alain. 2010. *Politics of Large Numbers: A History of Statistical Reasoning*. Cambridge, Massachusetts: Harvard University Press.

Dobbe, Roel, Sarah Dean, Thomas Gilbert, and Nitin Kohli. 2018. “A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics.”

Dwork, Cynthia, and Christina Ilvento. 2018. “Group Fairness Under Composition.” In *5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

Espeland, Wendy Nelson, and Stevens, Mitchell L. 1998. "Commensuration as a social process." *Annual Review Sociology* 24: 313–43.

Friedler, Sorelle A, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. “On the (Im) Possibility of Fairness.” *arXiv Preprint arXiv:1609.07236*.

Gajane, Pratik, and Mykola Pechenizkiy. 2018. “On Formalizing Fairness in Prediction with Machine Learning.” In *5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

Green, Ben. 2018. “‘Fair’ Risk Assessments: A Precarious Approach for Criminal Justice Reform.” In *5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

Hajer, Maarten A. 2006. “Doing discourse analysis: coalitions, practices, meaning.” In *Netherlands Geographical Studies*, edited by Margo van den Brink and Tamara Metze, 65–74.

High-Level Expert Group on Artificial Intelligence. 2019. “Ethics Guidelines for Trustworthy AI.” Brussels: European Commission.

Kearns, Michael, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. “Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness.” In *5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

Kitchin, Rob. 2014. *The Data Revolution*. London: Sage.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2016. “Inherent Trade-Offs in the Fair Determination of Risk Scores.” *arXiv Preprint arXiv:1609.05807*.

Latour, Bruno. 1983. “Give Me a Laboratory and I Will Raise the World.” In *Science Observed*, edited by K Knorr-Cetina and M. Mulkay, 141–70. Thousand Oaks: Sage.

Latour, Bruno, and Steve Woolgar. 1986. *Laboratory Life: The Construction of Scientific Facts*. Princeton, New Jersey: Princeton University Press.

Morozov, Evgeny. 2013. *To Save Everything, Click Here: The Folly of Technological Solutionism*. New York: Public Affairs.

Ogunyale, Tobi, De’Aira Bryant, and Ayanna Howard. 2018. “Does Removing Stereotype Priming Remove Bias? A Pilot Human-Robot Interaction Study.” *arXiv Preprint arXiv:1807.00948*.

Porter, Theodore M. 1995. *Trust in Numbers. The Pursuit of Objectivity in Science and Public Life.* Princeton: Princeton University Press.

Raff, Edward, and Jared Sylvester. 2018. “Gradient Reversal Against Discrimination.” In *5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

Selbst, Andrew D., Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. “Fairness and Abstraction in Sociotechnical Systems.” In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. FAT* ’19. New York, NY, USA: ACM. https://doi.org/10.1145/3287560.3287598.

Skirpan, Michael and Gorelick, Micha. 2017. "The authority of "fair" in machine learning." In *KDD'17 FATML Workshop*.

Wattenberg, Martin, Fernanda Viégas, and Moritz Hardt. 2016. “Attacking Discrimination with Smarter Machine Learning.” *Google Research* 17. https://research.google.com/bigpicture/attacking-discrimination-in-ml/.