Fair Enough? On (Avoiding) Bias in Data, Algorithms and Decisions

Francien Dechesne

Fair Enough? On (Avoiding) Bias in Data, Algorithms and Decisions

Francien Dechesne[1] [0000−0002−3511−9103]

eLaw Center for Law and Digital Technologies
Leiden University Law School
Steenschuur 25, 2311 ES Leiden, The Netherlands
f.dechesne@law.leidenuniv.nl

Abstract. This contribution explores bias in automated decision systems from a conceptual, (socio-)technical and normative perspective. In particular, it discusses the role of computational methods and mathematical models when striving for "fairness" of decisions involving such systems.

Keywords: Bias · Data analytics · Algorithmic Decision Systems · Fairness

1 Introduction

This contribution reflects on the discussion of fairness in so-called algorithmic decisions from a foundational and interdisciplinary perspective, based on a few of my favourite readings. It is a write-up of the keynote with the same title, delivered on the first day of the IFIP Summerschool. With the title, I have made a deliberate choice to highlight the following notions: bias, data, algorithms, and decisions. Data refers to the input as a possible point for intervention, preprocessing. Algorithms to the computational processing of the data. Decisions to the outcome - but as I hope to make clear: not just the outcome of the computational process.

If you are a scholar with a budding interest in the debate on fairness of algorithmic decision systems, you may search the internet for "What is bias?", and be surprised to find a tutorial titled: "All about Fabric Bias - Grain vs Bias and Cutting on the Bias."[1] Indeed, the third of four meanings on Dictionary.com explains this:

1. A particular tendency, trend, inclination, feeling, or opinion, especially one that is preconceived or unreasoned: illegal bias against older job applicants; the magazine's bias toward art rather than photography; our strong bias in favour of the idea.
2. Unreasonably hostile feelings or opinions about a social group; prejudice: accusations of racial bias.

[1] Cf. https://www.cucicucicoo.com/2017/02/fabric-bias-vs-grain/ - last accessed 3 Nov. 2019.

3. An oblique or diagonal line of direction, especially across a woven fabric.
4. Statistics. a systematic as opposed to a random distortion of a statistic as a result of sampling procedure.

What we can see in this definition is that essentially, bias is a deviation or inclination away from a certain standard, that can be either objective (as in meanings 3 and 4) or a matter of judgment (meanings 1 and 2). The relevant meanings of bias in the debate about so-called *algorithmic decisions* have to do both with systematic inclinations as in the fourth meaning, and with the judgments of meanings 1 and 2. Let us try to understand how they are connected.

2 Abstract Objectivity and Human Construction

There is a paradox in the way we talk about Artificial Intelligence, as something separate and somehow opposite from us. Doing so is attractive as it gives room for hope that it can deal with problems we cannot solve without us having to think about it. But at the same time, the autonomy that we attribute to it, feeds the fear that AI will soon figure out that human kind is the problem - so needs to be controlled. The fear, in short, that by granting AI too much autonomy, it will deprive us from ours.

We may recognise this fear in the discussion about so-called "algorithmic decision systems". Over the past half century we have both figured out that humans operate with bounded rationality (cf. the seminal work of [1]) and developed *universal machines* that operate strictly according to computational logics. So while we are bound to act according to biases and prejudices, which leads to suboptimal and/or unfair decisions, the machines are our hope to realise the modern ideal of rational actors. We hope the *artificial* part of AI can free us from our harmful irrationalities and prejudices.

"It is all human construction" said philosopher Maxim Februari in the 2017 Godwin lecture about the influence of digital technologies (in particular Big Data) on society [2]. Not only is AI a human creation, it is created to work in and for our human society, another human construction. The data we gather about behaviour, emotions, political views, cultural or ethnical groups, etc. are not merely things that are *given* to us (which is what *'data'* means in Latin); we constructed that social reality. Natural language is a great example of this, and we will get back to it later.

When algorithmic systems seem to be external to us, and objective, it is easy for us to believe that they -unlike us- have access to some external objective reality. This would make them suitable to be objective in their "decisions", without bias, so "objectively fair". The point here is that we have to be realistic however: algorithmic systems, the data we use in them, the interpretation of the outcome into a decision, are all in some way or another human/societal constructions, as is a normative notion like *fairness*. We should therefore look at questions of fairness through a socio-technical lens, acknowledging that the formal and computational gets meaning, including normative load, from the social context of application. But as we will see: formal and computational methods can also help in better understanding the problems around fairness, and contribute significantly to possible ways of dealing with them.

3 How is bias relevant in computational technologies?

Although not exactly the start, the Propublica discussion of the COMPAS recidivism scoring system [3] can be seen as a high profile catalyst for the debate on possible discriminatory effects of the application of data-analytics in all kinds of societal decision making. This case is interesting because the academic discussion of it has shown that mathematical tools may not *automatically* remove bias from our judgments. But it is particularly interesting because the discussion has demonstrated mathematically that different mathematical measures for fairness give different assessments [4] - and that some combinations are actually inconsistent, meaning that we cannot have them all.

Why would we even look to mathematics to help achieve fairness? Well, simply said, one could say that fair is treating equal cases equally, or similar cases similarly. This requires a theory that says when two cases can be considered the same, and mathematics works with abstract concepts like equivalence classes, or metric spaces which allow to define measures that indicate how close or far apart two different cases are in a certain sense. Such metric spaces play a central role in statistics and machine learning as well.

But still, while mathematics offers such tools, it will not tell you which measure to choose, which parameters are most important in determining the distance between cases, what characteristics are most salient to give a good representation of the issue in the abstract model you build of it. Neither does it tell you what "fair treatment" substantively means. Fairness is a normative concept. We can abstract it into maths, but only after deciding what type of "equal treatment" we find most appropriate. For example (see Figure 1), when dividing limited resources, do give each individual the same amount (equality)? Or do we take into account what people actually need to achieve the same outcome (equity)? But we could also look for a way to reshape the situation so that the problem disappears. In the picture, for example: what if we removed the fence?

What I think the picture in Figure 1 nicely illustrates, is Melvin Kranzberg's famous quote in the presidential address to the Society for the History of Technology [6]:

> "Technology is neither good not bad, nor is it neutral."

It points to the inherent interaction between societal constructions, and technical ones. How is technology not neutral? To understand this, I find Don Ihde's notion of technological *mediation* a useful one: we perceive and construct our understanding of the world partially through the tools that we have at hand. It changes how we perceive the world as it *is*, as it *can be* (what we consider to be possible) and also: as it *should be* (what we consider to be desirable or necessary). [7]

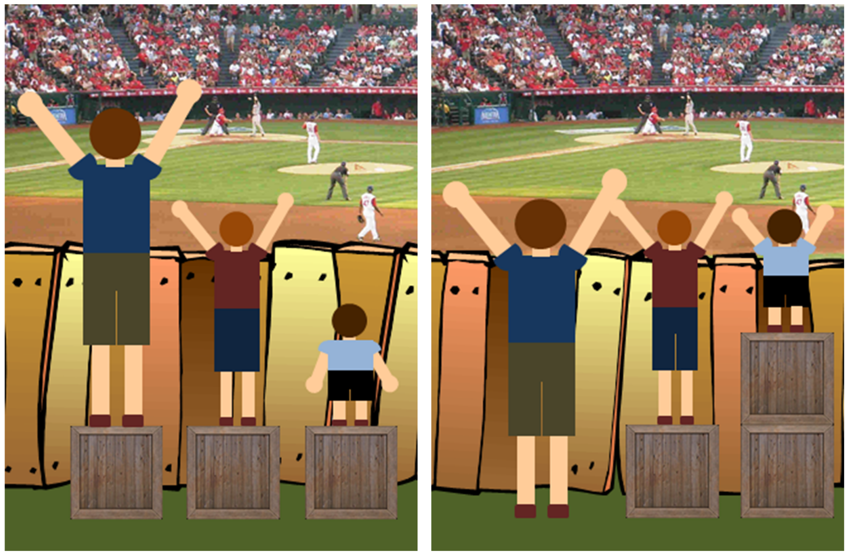

Fig. 1. A visualisation of two arguably fair, but incompatible ways of distributions of limited resources: equality (left) and equity (right). This visualisation went viral: [5].

If we go back and think about AI in this way, it influences how we think about what is natural, what is real, what is human, what is fair…etc. This in turn influences how we design solutions for what we see as problems. For the particular case of "algorithmic decisions" I would like to recommend two books that elaborate how this transformation unfolds of how we perceive the world, what we see as possible and what we think should be the case.

The first one is legal scholar Bernhard E. Harcourt's *Against Prediction* [8]. Rather than speaking in terms of AI or algorithms, it discusses the effect of relying on statistical methods, correlations, in the judicial practice and the idea these methods are effective in preventing crime. In the book, Harcourt develops three main critiques. Firstly, the over-reliance on rational action theory in claims about the effectivity of these methods in reducing crime. Secondly, the so-called *ratchet effects*: disproportionate policing on situations that come out of the predictions as *high-risk* will increase statistical bias, and may effectively increase actual crime by leaving other areas under-policed (and crime there under-registered). It is important to also acknowledge that such practices may have an impact on the community through the transformations of the social meaning of police conduct, which affect social norms on priorities in law enforcement. In that way, thirdly, technical advances implicitly reshape the justice system.

Another book comes from a mathematician, who, working in the commercial practice of data-analytics, saw how mathematical tools can be applied with fundamental societal impact: Cathy O'Neil's *Weapons of Math Destruction* [9]. In the book she identifies three main issues with data-driven (risk) scoring mechanisms: the feedback loops they establish in whatever they measure (cf. Harcourt's *ratchet effects*), their scale of application across contexts, and their opacity. The opacity not only comes from the fact that mathematical methods feel impenetrable (and unobjectionable) to non-mathematicians, but also because applications of machine learning are *inherently* opaque. And on top of that, there is the opacity that comes with the fact that many systems that are used, also in public institutions, are secret under trade secrets protection laws.

O'Neil makes an important observation regarding the source of bias in the use of algorithmic systems, and that is that the definition of what constitutes success for the system is where biases originate. This steers the ship in a certain direction (such as efficiency, profitability, cost minimisation etc.) and may severely compromise other goals or (non-functional) requirements of the system.

4 Biases in natural language

As indicated before, natural language and natural language processing demonstrate interestingly how bias, machine learning and the remark that "it's all human construction" come together. As we know from Google search suggestions and Google Translate, machine learning techniques are very successful in predicting what we might search for, or translating words and phrases between languages. These systems learn from huge amounts of textual data that reflect how we actually use language.

But such techniques then also absorb tendencies of usage that are not predefined in the dictionary meanings of words. In Google search suggestions, the type of suggestions that popped up if you would type in the search bar: "men are…", would be surprisingly (and sometimes offensively) different from the suggestions for "women are…" [10]. Similar effects could be noticed in translations, for example when translating back and forth from Turkish, which has gender neutral pronouns, to English [11].

The phenomenon of biases in our use of natural language has been studied using machine learning techniques such as Word2Vec. Tolga Bolukbasi et al.'s 2016 paper with the telling title: "Man is to Computer Programmer as Woman is to Home-maker" [12] uses this technique to show implicit gender biases in our use of words that are gender neutral in principle. The paper also proposes ways of removing such biases by identifying such gender neutral words and performing an intervention on their vector representations as induced from the texts (giving such words the same distance to "he" as to "she"). While this debiases the representation of our natural language use, it of course does not debias our use of natural language itself, nor the social, cultural and historical structures and meanings that it encodes (as [13] in fact demonstrates).

One could question how fundamentally effective and meaningful this is, and the authors of the paper initiate this discussion themselves. But in any case, for the Google Search and Translate examples mentioned above, Google has in the meantime indeed intervened. You will notice that certain queries of the form "[group of people] are..." no longer generate any suggestions, and also the identified issue with translating back and forth to gender neutral languages has been acted upon [15].

5 Should we do something about bias?

At this point it is maybe good to take a step back and revisit the concept of *bias*. While the word bias in everyday language (and most writing about algorithmic decision systems) is primarily used to indicate (unjustified and/or harmful) prejudices, remember that it in essence refers to an inclination away from a certain standard: e.g. bias in statistics is about structural misrepresentation, and bias in behavioural decision theory is about tendencies that deviate from the *rational* choice. In principle bias does not have to have a negative normative load, but calling something a bias does imply an assumed norm or standard (e.g. rational choice).

Deviation from a norm depends on the norm and the measure, and whether bias is a bad thing depends on the normative context. Because many of the examples about bias and algorithmic systems touch upon discrimination, it is also important to highlight that discrimination (in those discourses) is legally codified. Discrimination as a legal concept is introduced against harmful distinctions between groups of people, defined in law in terms of certain sensitive characteristics (such as race, gender, religion etc.). Not only intended, explicit differential treatment (in US law: *disparate treatment*) counts as discrimination, but also implicit harmful differences with effect on certain groups of people (*disparate impact*). An excellent overview of how algorithmic systems may fit legal definitions of discrimination in the latter sense can be found in the paper "Big Data's Disparate Impact" by Solon Barocas and Andrew Selbst [14].

Bias in algorithmic systems enters at several layers, and it is impossible to completely disentangle the role of different sources of bias. At the level of the **data**: they may be incorrect, incomplete, or non-representative in a certain systematic way. Bias in data is introduced in what we can efficiently and accurately measure, how good our proxies work and what parameters we think are significant - for the definition of success, which in turn is determined by one stakeholder, and interpreted and validated by others. When training our systems of machine learning **algorithms**, bias is introduced by the availability and our selection of the training data, the choice of the algorithm, the choice what to optimize for in the learning process and what error margins are considered to be acceptable. This in turn is closely related to how we interpret and apply the outcome of such a system. While we may refer to the outcome of the computation as the (algorithmic) **decision**, in essence, the number is just that: the outcome of a (symbolic) computation. It is our (human) interpretation of that outcome, and the translation of that interpretation into a judgment or action (or our ascription of an intention to that outcome), that elevates the outcome of a computation into a decision. Here I would like to refer back to Kranzberg's quote that technology - including the *application* of computational techniques - is not neutral. It mediates and co-shapes our ideas of what a decision is - and also what a *good* decision is.

Through algorithmic systems, that are very powerful in highlighting patterns in data, we may attribute to those systems the powers to overcome human biases. But we have to realise that our biases seep into these systems at all layers, as trade-offs need to be made and will be made by people who can not oversee all eventual uses of the system - so they have to act on certain preconceptions. For example, underlying the claims of fairness of the recidivism scoring algorithm Compas ([3], see above), there was an "assumption that the algorithm's predictions were inherently better than human ones" [16]. It can actually be shown that in this case the system's predictions do not outperform human judgments - at least not in accuracy.

6 Formal methods as tools for understanding

What do these reflections mean for what we can do to deal with harmful effects of biases in algorithmic systems? By now, different papers have appeared presenting different (technical) interventions to address bias at the level of data and algorithms [17–19]. Also, the first comparative studies appear, finding that a large number of measures are essentially similar but just make different trade-offs. It turns out there is a very high dependency on the way the data are preprocessed [20].

The outcomes of the systems only have meaning in their (historical, societal, cultural, domain-specific) context, and so does what counts as fair. This points at the fact that there can be no mathematical solution bringing us universal and objective fairness. Trade offs need to be made at all points, and the models that are trained to represent some physical or conceptual reality, can only be representations of a part of that reality - and it can at its best be made to be accurate for another part of that.

While mathematical methods cannot provide universal fairness solutions, they *can* be used to clearly point us at their limitations - which is useful knowledge in the light of underlying assumptions about some kind of technological superiority. It will help us understand in which ways the systems can improve human decision making, but also how this improvement depends on an implicit mapping between "reality" and its symbolic representation within the data and the algorithmic system. In the paper "The (im)possibility of fairness" by Friedler, Venkatasubramanian and Scheidegger [21], the authors use formal methods to pull apart the different *spaces* involved: the (intractable) construct space (the aspect you would like to measure), the observed space (the proxy for the construct, something measurable, the inputs for the decision system) and the decision space (the outputs of the system). They use this formal model of algorithmic decisions to mathematically prove that certain different notions of fairness can only be realised together under extreme, implausible metaphysical assumptions. This shows that it is really necessary to make normative choices of what is the most appropriate measure of fairness for a given situation.

Another paper, by Zliobaite and Custers [22], demonstrates mathematically how bias mitigating measures taken at different layers may interact - or rather: counteract. Data protection, as for example codified in the European Union's General Data Protection Regulation (GDPR [23]), has data minimisation as one of its principles in order to protect data subjects. In particular, Art. 9 of the GDPR *in principle* prohibits the processing of special categories of personal data ('sensitive data') such as data revealing racial or ethnic origin or religious beliefs. But if we leave out those attributes to prevent disparate treatment, this also means that we make it impossible to apply certain automated methods to control for harmful biases inherited in the data.

While law is often accused of being too slow to adapt to changing circumstances, one could say that mathematical laws in fact do not move at all. The application of any mathematical formalism comes with underlying assumptions on the structure of the problem it models and addresses. It thereby has a tendency to abstract away at least some forms of change or feedback loops of the use of the model, that may over time lead to violation of those underlying assumptions. There is also mathematical work demonstrating this for fairness measures [24].

7 Concluding remarks

In this contribution, I have reflected on a few years of attention from the mathematical and computational sciences to bias in data, algorithms and decisions. Whereas one might expect neutrality and objectivity from computational systems, I have discussed how taking actual decisions involves human interpretations and judgments - in the data sets the system works with, the design of the algorithmic system, and turning the outcome into a decision. Assumptions, selections and priority judgments are made, on the basis of availability, efficiency, or other reasons, and through all of these, biases enter the picture.

Bias in its most neutral description is some systematic deviation from a standard. To which extent bias can be harmful or unfair, depends on the circumstances. Fairness in decisions based on machine learning is inherently sociotechnical and context sensitive. Formalisation of the problem will not provide us with universal solutions for fairness in real life, but the good news is: mathematical methods do bring us some insights in where the limitations lie both in formal approaches and in real life solutions to unfairness. It can point us to the necessity of looking at the temporal dimension as well, to capture feedback loops and historical factors.

Different fairness measures may be provably inconsistent in most realistic cases. Working with different measures and formally comparing them is very useful, in particular for analysing the problems and deciding which measure is most fitting where. Ultimately, whether a system is fair enough depends on whether the trade-offs we embed in the system, implicitly or explicitly, are right for the societal context of application. And what is right is subject to constant societal re-evaluation. This may also lead to the conclusion that for certain types of decisions, in the interest of fairness and justice, algorithmic systems are not the right tools.

Acknowledgment: This work is part of the SCALES project funded by the Dutch Research Council NWO MVI-program under project number 313-99-315.

References

1. Daniel Kahneman, Paul Slovic, Paul, Amos Tversky. Judgment under Uncertainty: Heuristics and Biases. Cambridge, UK: Cambridge University Press. (1982)
2. Maxim Februari, “Het is allemaal mensenwerk” (in Dutch). Godwin lecture, 5 May 2017. De Correspondent. https://decorrespondent.nl/6692/de-datahonger-van-staten-en-bedrijven-zet-veel-meer-op-het-spel-dan-uw-privacy-alleen/1514846110316-d1a5748d
3. Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, Machine Bias - There’s software used across the country to predict future criminals. And it’s biased against blacks. ProPublica May 23, 2016. Available online: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
4. Arvind Narayanan, 21 Definitions of Fairness. Tutorial at the FAT* conference, February 2018, New York. https://www.youtube.com/watch?v=jIXIuYdnyyk
5. Craig Froehle, The Evolution of an Accidental Meme. How one little graphic became shared and adapted by millions. Medium. April 14, 2016. https://medium.com/@CRA1G/the-evolution-of-an-accidental-meme-ddc4e139e0e4
6. Melvin Kranzberg, Technology and History: “Kranzberg’s Laws”. Technology and Culture, 27(3), 544-560. (1986) doi:10.2307/3105385
7. Peter-Paul Verbeek, Animation: Explaining Technological Mediation. June 2017. Available at https://vimeo.com/221545135.
8. Bernard Harcourt, Against Prediction. University of Chicago Press. (2006)
9. Cathy O’Neil, Weapons of Math Destruction - How big data increases inequality and threatens democracy. Crown (2016)
10. Issy Lapowsky, Google Autocomplete Still Makes Vile Suggestions. Wired, December 2, 2018. Available at https://www.wired.com/story/google-autocomplete-vile-suggestions/.
11. Nikhil Sonnad, Google Translate’s gender bias pairs “he” with “hardworking” and “she” with lazy, and other examples. Quartz, 29 November 2017. Available at https://qz.com/1141122/google-translates-gender-bias-pairs-he-with-hardworking-and-she-with-lazy-and-other-examples/
12. Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Advances in Neural Information Processing Systems, pages 4349-4357, 2016.
13. Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. Science 356.6334 (2017): 183-186.

14. Solon Barocas and Andrew D. Selbst, Big Data's Disparate Impact.California Law Review104 (2016).
15. James Kuczmarski, Reducing gender bias in Google Translate. Google Blog, December 2, 2018. Available at https://www.blog.google/products/translate/reducing-gender-bias-google-translate/.
16. Ed Yong, A Popular Algorithm Is No Better at Predicting Crimes Than Random People. The Atlantic, January 17, 2018. Available at https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/
17. Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In Proceedings of the 26th International Conference on World Wide Web (WWW '17). 1171-1180. DOI: https://doi.org/10.1145/3038912.3052660
18. Sahil Verma and Julia Rubin. 2018. Fairness Definitions Explained. In FairWare'18: IEEE/ACM International Workshop on Software Fairness, May 29, 2018, Gothenburg, Sweden. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3194770.3194776
19. Indre Zliobaite, 2015. A survey on measuring indirect discrimination in machine learning. arXiv:1511.00148 [cs.CY]
20. Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). ACM, New York, NY, USA, 329-338. DOI: https://doi.org/10.1145/3287560.3287589
21. Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian. 2016. On the (im)possibility of fairness. arXiv:1609.07236 [cs.CY]
22. Zliobaite, Indre and Custers, Bart, Using Sensitive Personal Data May Be Necessary for Avoiding Discrimination in Data-Driven Decision Models. Artificial Intelligence and Law (24): 183-201. Available at SSRN: https://ssrn.com/abstract=3047233
23. European Council. General Data Protection Regulation. Regulation (EU) 2016/679
24. Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, Moritz Hardt. Delayed Impact of Fair Machine Learning. Proceedings of the 35th International Conference on Machine Learning, PMLR 80:3150-3158, 2018. See also the blog-post at https://bair.berkeley.edu/blog/2018/05/17/delayed-impact/