Rediscovering the human in AI design for fairness[1]

Authors: Ana Carolina de Assis Nunes[2], and Shaozeng Zhang[3]

Abstract:

This paper is an initial report of our fair AI design project by a small research team made up of anthropologists and computer scientists. Our collaborative project was developed in response to the recent debates on AI's ethical and social issues (Elish and boyd 2018). We share this understanding that "numbers don't speak for themselves," but data enters into research projects already "fully cooked" (D'Ignazio and Klein 2020). Therefore, we take an anthropological approach to observing, recording, understanding, and reflecting upon the process of machine learning algorithm design from the first steps of choosing and coding datasets for training and building algorithms. We tease apart the encoding of social-cultural paradigms in the generation and use of datasets in algorithm design and testing. By doing so, we rediscover the human in data to challenge the methodological and social assumptions in data use and then to adjust the model and parameters of our algorithms. This paper centers on tracing the social trajectory of the Correctional Offender Management Profiling for Alternative Sanctions, known as the COMPAS dataset. This dataset contains data of over 10,000 criminal defendants in Broward County in Florida, the U.S. Since its publication, it has become a benchmark dataset in the study of algorithmic fairness and was also used to design and train our algorithm for recidivism prediction. This paper presents our observation that data results from a complex set of social, political, and historical assumptions and circumstances and demonstrates how the social trajectory of data can be taken into the design of AI as automated systems become more intricate into our daily lives."

Key words:

Fairness, machine learning, human in the loop, social trajectory of dataset, data biography

---

[1] de Assis Nunes, Ana Carolina, and Shaozeng Zhang. Oct. 2021. "Rediscovering the human in AI design for fairness." Annual conference of the Society for the Social Studies of Science (4S), Toronto, Canada. Presentation of this paper at the 4S conference is accessible at https://media.oregonstate.edu/media/t/1_2btx97hh . This paper is based on the the research project "Human-in-the-Loop Fairness Optimization in Machine Learning with Minimax Loss and an Abstain Option" funded by the U.S. National Science Foundation (NSF) AI-DCL EAGER grant.
A general introduction of this project is accessible at https://blogs.oregonstate.edu/designingfairai/.

[2] Oregon State University, nunesa@oregonstate.edu
[3] Oregon State University

Introduction

Digital technologies and artificial intelligence are no longer novel topics. Those words permeate every bit of our existence, and their uses regulate several aspects of our lives. Kate Crawford (2021) writes, "Artificial intelligence is not an objective, universal, or neutral computational technique that makes determinations without human direction." For this author, its systems are embedded in social, political, cultural, and economic worlds, shaped by humans, institutions, and imperatives that determine what they do and how they do it. They are designed to discriminate, amplify hierarchies, and encode narrow classifications.

Crawford highlights that when applied in social contexts such as policing, the court system, health care, and education, they can reproduce, optimize, and amplify existing structural inequalities as AI systems are expressions of power.

In this scenario, with AI scoring, it's no different, and Crawford (2021) affirms that's what happens when AI enters traditional domains of state logic such as with law enforcement and border control. The author asks, "How can we intervene to address interdependent issues of social, economic, and climate injustice? Where does technology serve that vision? And are there places where AI should not be used, where it undermines justice?"

With these questions in mind and knowing that such technologies tend to increase and not decrease in upcoming years, our project touches on some of these premises.

Situating AI

AI for predicting social outcomes is highly controversial. According to Holton and Boyd (2021:182) five elements are central to the definition of artificial intelligence. They are: big data, algorithms, machine learning, sensing and logic/rationale. In this presentation we're focusing on one of these characteristics, which is big data. Big data, which has been called the "new oil" (Ray K., Strasser 2020), is in essence a large dataset, such as Compas, with a few other characteristics.

Elish and boyd (2018) write that big data was "born of big business." According to them, big data has been defined by the 3Vs: volume, velocity and variety (59) which can be misleading as Elish and boyd wrote. Still, challenging this perspective, Kitchin and McArdle (2016) after analyzing 26 datasets, have actually found that besides the 3Vs, exhaustivity, resolution,

indexicality, relationality, extensionality and scalability are also important for this definition, being velocity and exhaustivity more important than the 3Vs. The question of how the huge amounts of data that end up in datasets such as Compas are collected are also important (Birhani 2021; O'Neill 2016; D'Ignazio & Klein 2020 and others) especially in a data-saturated world (Knox 2018).

Abeba Birhani (2021) has been very critical of the wave of research on the field commonly referred to as "algorithmic fairness," and if discussions on this topic are becoming a common place in the humanities and social scientists are eager to show where AI fails; this discussion, however, is not so advanced in many computer sciences departments. Birhani (2021) highlights that the solutions proposed by many researchers in this area such as "fine-tuning specific models, making datasets more inclusive and the idea of de-biasing datasets do not address the wider picture. These solutions, actually, put forward technical fixes and do not center individuals and communities disproportionately affected by these technologies (2). Birhani calls for efforts in ethical AI to center the material condition and concrete consequences an algorithm tool is likely to bring (2021:6). So, in a way, our project was also a way to bring this discussion closer to computer scientists, sometimes divorced from such topics.

David Moats and Nick Seaver (2019) have written on the difficulty of leading multidisciplinary projects with mixed team of computer scientists and anthropologists in their paper "You social scientist love mind games": experimenting in the "divide" between data science and critical algorithm studies, mentioning that "part of the divide between data scientists and their qualitative critics has to do with subtle differences between how the two camps (and divisions within those two camps) become accountable to each other" (p. 8). Citing Paul Dourish, Seaver (2017:2) has written that if we want to understand engineers and get them to listen to us, we need to use terms as they do. This is also part of what we tried to do in our research.

Doing research in such a mixed team has never been easy. Anthropologist Diana Forsythe shared her experience in working with artificial intelligence designers in the 1980s and 1990s, with some of her collaborators referring to her work as sometimes "new, soft, and unscientific" (2001:133). In our case, the conversation sometimes didn't seem to happen at ease. And even later, when conducting focus groups, it sometimes felt like we were talking different languages. In this paper, we tease apart the encoding of social-cultural paradigms in the generation and use of datasets in algorithm design and testing. By doing so, we rediscover the human in data to

challenge the methodological and social assumptions in data use and to adjust the model and parameters of our algorithms.

Literature review

Consider context is one of the six principles of Data Feminism expressed in the book of the same name by Catherine D'Ignazio and Lauren F. Klein (2020)—among other things, the chapter discusses the importance of not taking numbers at face value, but instead, considering the context of data production. According to the authors, we may not believe in 1) the title of the database 2) the documentation 3) the marketing hype (p.152). The "numbers speak for themselves" narrative is a critique to the premise that data are a raw input, instead, write D'Ignazio and Klein, data enter into research projects already "fully cooked", what means: data is the result of a complex set of social, political and historical circumstances (p. 159). According to the authors: Instead of taking data at face value and looking toward future insights, data scientists can first interrogate the context, limitations and validity of the data under use. (…) to consider the cooking process that produces "raw" data. (…) exploring and analyzing what is missing from a dataset is a powerful way to gain insight into the "cooking" process—of both the data and of the phenomenon it purports to represent (p. 160).

This idea of "fully cooked" data is also shared by Crystal Biruk, who in her 2018 book Cooking Data writes about the social lives of numbers, rather than viewing them as stable objects and measures of reality. For this author, data are units of information—such as a number, response, or code written into a box on a survey page by a data collector; the author also shows through her research that data is always cooked by the processes and practices of production, at the same time that she seeks to destabilize the binaries surrounding all sorts of data.

Paul Dourish and Edgar Gómez Cruz (2018:8), also writing from the perspective that data do not speak for themselves, highlight that "data makes sense only to the extent that we have frames for making sense of it, and the difference between a productive data analysis and a random-number generator is a narrative account of the meaningfulness of their outputs." The author positions anthropologists and ethnographic researchers as especially apt to study the narratives told with and through data, as well as the possibilities and limits of data analysis and its social contexts. Following these authors, the approach we take in this initial report-research is ethnographic not

in the sense that it has arisen through an ethnographic investigation but rather in that it is informed by an ethnographic outlook, or a narrative approach to data practices (2018:8.). These same authors (2018:6), writing about stories told through and with data, acknowledge that data narratives about data help to ‘‘fix’’ data temporally. That is, the accounts that data narratives offer are ones that make sense of data within an evolving context, and so stabilize it in the sense that they situate it within a landscape of recognizable objects." While Deborah Lupton (2018:9) writes that “personal data, like other forms of mediated representations of bodies and selves, are dynamic assemblages of humans and nonhumans that are constantly subject to change.” These perspectives are important in questioning the kind of data feeding recidivism algorithms, or AI for predicting social outcomes, what computer scientist Arvind Narayanan (2021) has called “AI snake oil”. M. C. Elish and dana boyd (2018) have also written about the dubious promises of AI technology, which is sometimes defined and thought about more in terms of marketing than of what it actually does. It’s a question of marketing hype for the former and the latter.

Discussion

Sarkar, Yang and Vihinen (2020) write that benchmark datasets can be used for method training and testing. According to them, high-quality benchmark datasets are valuable and difficult to generate. These authors come up with a series of characteristics to define a benchmark dataset. For them, the principal characteristics of benchmark datasets are its relevance or capacity the dataset must have to capture the characteristics of the investigated property. Its representativeness, or the implication that the dataset should be of sufficient size to allow statistical studies but may not need to include all known instances. Its non-redundancy, meaning it should exclude overlapping cases within each dataset. Within this criterion, benchmark datasets should also contain experimentally verified cases meaning that the method performance comparisons have to be based on experimental data; Positive and negative cases. Comprehensive assessment should be based both on positive (showing the investigated feature) and negative (not having effect) cases; Scalability, so that it should be possible to test systems of different sizes. As well as reusability. As datasets are expensive to generate, meaning it should have similar applications or usage in new areas (Sarkar, Yang and Vihinen 2020:2).

Regardless of this technical definition, J. Buolamwini and T. Gebru highlights that "a demographic group that is underrepresented in benchmark datasets can nonetheless be subjected to frequent targeting" and concludes that other academic works "should explore gender classification on an inclusive benchmark" (2018:2-12). That's a little bit of what happens with the Compas dataset.

Compas dataset

For this part of the text, we followed Heather Krause's Data biography's approach to the study of the Compas dataset. Our perspective is also informed by what Paul Dourish and Gómez Cruz (2018) calls "ethnographic outlook" in the study of algorithms. Krause's approach includes asking the following questions about a dataset:

1. Where did it come from?
2. Who collected it?
3. When?
4. How was it collected?
5. Why was it collected?

Compas is an acronym for Correctional Offender Management Profiling for Alternative Sanctions, is a case management and decision support tool used by U.S. courts to assess the likelihood of a defendant becoming a recidivist.[2] The tool was developed in the 1990s, by a company called Northpointe, Inc. (now Equivant) which set out to create what is now known as Compas, a statistically based algorithm designed to assess the risk that a given defendant will commit a crime after release. In 2012, after years of development, the state of Wisconsin implemented Compas into its state sentencing procedures, at which point Compas assessments officially became a part of a defendant's presentence investigation (PSI) report.[3]

Compas's algorithm uses a variety of factors, including a defendant's own responses to a questionnaire, to generate a recidivism-risk score between 1 and 10. In general terms, this is accomplished by comparing an individual's attributes and qualities to those of known high-risk offenders. Based on this score, Compas classifies the risk of recidivism as low-risk (1 to 4), medium-risk (5 to 7), or high-risk (8 to 10). This score is then included in a defendant's PSI report supplied to the sentencing judge. As a result, a defendant's sentence is determined—to at least some degree—by Compa's recidivism risk assessment.[4]

According to researchers Cynthia Rudin, Caroline Wang, and Beau Coker "Compas analysis is complicated. It is based on up to 137 variables (Northpointe, 2009) that are collected from a questionnaire." According to them, "this is a serious problem because typographical or data entry errors, data integration errors, missing data, and other types of errors abound when relying on manually entered data. Individuals with long criminal histories are sometimes given low Compas scores (which labels them as low risk), and vice versa." For these authors, "a separate issue with Compas is that it is proprietary, which means its calculations cannot be double-checked for individual cases, and its methodology cannot be verified. Furthermore, it is unclear how the data Compas collects contribute to its automated assessments.

Denton et al. (2021) unpacked some aspects of the ImageNet dataset in their article *On the genealogy of machine learning datasets: A critical history of ImageNet*. The authors conceptualize ML datasets as a type of informational infrastructure and motivate genealogy as a method of examining the histories and modes of constitution of ML datasets (p. 11). The authors suggest that to understand how and why ML systems fail marginalized communities, we need to write critical histories of our present in which ML datasets are understood both as infrastructural and genealogical objects of inquiry (p. 2)." For them, datasets have a historical and temporal dimension, and are situated artifacts (p. 2). They cite Latour to highlight how the more naturalized ML datasets become, the more likely they are to be treated as value-neutral scientific artifacts and unquestioningly adopted by ML practitioners. In this spirit, we ask, how did the Compas dataset come to matter in machine learning research involving the discussion of fairness?

To which we add, Compas received some visibility and scrutiny after ProPublica's 2016 study, in which it attempted to reconstruct Compas methodology[4]. The tool known as Compas and the dataset associated with it, were analyzed, and scrutinized by ProPublica which obtained two years' worth of Compas scores from the Broward County Sheriff's Office in Florida. They received data for all 18,610 people who were scored in 2013 and 2014. ProPublica then looked at more than 10,000 criminal defendants in Broward County, Florida, and compared their predicted recidivism rates with the rate that occurred over a two-year period and matched the criminal records to the Compas records using a person's first and last names and date of birth. This is the same technique used in the Broward County Compas validation study conducted by researchers

[4] https://towardsdatascience.com/compas-case-study-fairness-of-a-machine-learning-model-f0f804108751

at Florida State University in 2010. They downloaded around 80,000 criminal records from the Broward County Clerk's Office website[5]. Because Broward County primarily uses the score to determine whether to release or detain a defendant before his or her trial, they discarded scores that were assessed at parole, probation, or other stages in the criminal justice system. That left them with 11,757 people who were assessed at the pretrial stage. It's important to highlight that the Compas tool assigns defendants scores from 1 to 10 that indicate how likely they are to reoffend based on more than 100 factors, including age, sex and criminal history. Notably, race is not used. So, to determine race, ProPublica used the race classifications used by the Broward County Sheriff's Office, which identifies defendants as black, white, Hispanic, Asian and Native American. In 343 cases, the race was marked as Other.

It's important to mention that the dataset we used in this research is the one provided by ProPublica and not by Northpointe. Propublica's dataset has 10 features compared with 137 from NorthPointe. ProPublica's model was 61% effective, and they found a racial bias in favor of white people since black people were more likely to be falsely flagged as a high risk for recidivism. Northpointe criticized ProPublica's research and the results it generated. But since 2016 and after ProPublica's public attention on this topic, other researchers started using the Compas dataset to propose other validations on the algorithm. That's when the Compas dataset came to matter. Bao et al. (2021) write about how Risk Assessment Instruments (RAI) datasets are commonly used in algorithmic fairness research due to benchmarking practices of comparing algorithms on datasets used in prior work. The author criticizes benchmark datasets on fairness because they're a generic real-world example; these datasets should also be avoided according to them due to the different laws, practices, and data acquisition methods available in different states (Bao et al. 2021:8). But here we follow Krause's suggestion.

Conclusion

In this paper, we offered a brief overview of research in AI for social outcomes. We explored more closely the Compas dataset, used by our small research group, as a way to show how numbers don't speak for themselves, but that data enters the research already cooked. We also briefly mentioned that while discussions about the pitfalls of AI research for social outcomes are

[5] https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

very advanced in the humanities and social sciences, the same is not true in some computer science circles, in a way that our work became also one of informing our research colleagues about the background and challenges of working with Compas dataset, what reiterate points already discussed by Seaver (2017) and Dourish (2018).

Using the approach of data biography was an important way to situate the data our team used, and highlighting that the numbers were not only numbers, but that they referred to human beings. Bringing these particularities about COMPAS dataset illustrates the importance of not taking numbers at face value, but instead, considering the context of data production. Instead of taking data at face value and looking toward future insights, data scientists can first interrogate the context, limitations and validity of the data under use.

References

Bao, M., Zhou, A., Zottola, S., Brubach, B., Desmarais, S., Horowitz, A., Lum, K., & Venkatasubramanian, S. (2021). It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks. *ArXiv:2106.05498 [Cs]*.

Birhane, A. (2021a). Algorithmic injustice: A relational ethics approach. *Patterns*, *2*(2), 100205.

Birhane, A. (2021b). The Impossibility of Automating Ambiguity. *Artificial Life*, *27*(1), 44–61.

Biruk, C. (2018). *Cooking Data: culture and politics in an African research world*. Duke University Press.

Buolamwini, J., & Gebru, T. (n.d.). *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*. 15.

Crawford, K. (2021). *The Atlas of AI*. Yale University Press.

D'ignazio, C., & Klein, L. F. (2020). *Data feminism*. MIT press.

Denton, E., Hanna, A., Amironesei, R., Smart, A., & Nicole, H. (2021). On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society*, *8*(2), 205395172110359.

Dourish, P., & Gómez Cruz, E. (2018). Datafication and data fiction: Narrating data and narrating with data. *Big Data & Society*, *5*(2), 2053951718784083.

Dourish, P. (2016). Algorithms and their others: Algorithmic culture in context. *Big Data & Society*, *3*(2), 2053951716665128.

Elish, M. C., & Boyd, D. (2018). Situating methods in the magic of Big Data and AI. *Communication monographs*, *85*(1), 57-80.

Forsythe, D. (2001). *Studying those who study us: An anthropologist in the world of artificial intelligence*. Stanford University Press.

Holton, R., & Boyd, R. (2021). 'Where are the people? What are they doing? Why are they doing it?'(Mindell) Situating artificial intelligence within a socio-technical framework. *Journal of Sociology*, *57*(2), 179–195.

Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, *3*(1), 2053951716631130.

Lupton, D. (2018). How do data come to matter? Living and becoming with personal data. *Big Data & Society*, *5*(2), 2053951718786314.

Moats, D., & Seaver, N. (2019). "You Social Scientists Love Mind Games": Experimenting in the "divide" between data science and critical algorithm studies. *Big Data & Society*, *6*(1), 2053951719833404.

Sarkar, A., Yang, Y., & Vihinen, M. (2020). *Variation benchmark datasets: Update, criteria, quality and applications*. *2020*, 16.

Seaver, N. (2017). Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society*, *4*(2), 205395171773810.