What does "fairness" mean for machine learning systems?

"Fairness" is a ubiquitous term in the artificial intelligence (AI) and machine learning (ML) space. Most principles for responsible and ethical AI include "fairness". But what does that mean in practice and what is a "fair" ML system? This brief explores "fairness" broadly, then dives into the default fairness approach in ML and associated challenges. It ends with tools and considerations for those developing, managing and using ML systems.

I. Understanding "fairness"

Fairness is a confusing concept. Fairness is commonly defined as the quality or state of being fair, especially fair or impartial treatment. But what's fair can mean different things in different contexts to different people.[1]

Fairness has different definitions across disciplines too. This is captured in a paper by Deirdre Mulligan, Joshua Kroll, Nitin Kohli and Richmond Wong.[2] For example:

Law: fairness includes protecting individuals and groups from discrimination or mistreatment with a focus on prohibiting behaviors, biases and basing decisions on certain protected factors or social group categories.

Social science: "often considers fairness in light of social relationships, power dynamics, institutions and markets."[3] Members of certain groups (or identities) that tend to experience advantages.

Quantitative fields (i.e. math, computer science, statistics, economics): questions of fairness are seen as mathematical problems. Fairness tends to match to some sort of criteria, such as equal or equitable allocation, representation, or error rates, for a particular task or problem.

Philosophy: ideas of fairness "rest on a sense that what is fair is also what is morally right."[4] Political philosophy connects fairness to notions of justice and equity.

Even within disciplines, definitions can differ. It's no wonder then that fairness in machine learning systems has caused confusion.

II. The default fairness approach in machine learning & its issues

ML researchers and practitioners tend to use a quantitative perspective as the primary lens for fairness. They focus on constructing an optimal ML model subject to fairness constraints (a "constrained optimization problem"). The constraints the model is subjected to can be informed from law, social science and philosophy perspectives.

Commonly, constraints tend to be around sensitive, legally protected attributes. ML researchers and practitioners want the model to perform as optimally as possible while

also treating people "fairly" with respect to these sensitive attributes. Fairness can be defined at the individual level (such as ensuring that similar individuals are treated similarly) or at the group level. In the latter case, this is done by grouping people into categories and ensuring that the groups are treated somewhat equitably. While fairness for a group can be formulated in different ways, the simplest is pursuing demographic parity across different subgroups (meaning each subgroup receives the positive outcome at equal rates / the same proportion). With demographic parity, membership in a protected class should have no correlation with the decision.

This quantitative approach can be problematic. Approaches tend to be narrowly specified and don't always capture the nuances and various conceptions of fairness. Pursuing demographic parity in particular may seem like a good solution but is a simplistic approach to fairness that can still be at odds with other definitions of fairness[5] -- such as justice. Also, even if satisfying demographic parity based on gender, for example, when overlaying race on top of gender, this parity can be off. It's also important to not only look at parity in terms of how ML systems allocate resources, but also in how they choose not to allocate resources. See Box 1 for an example to make this more clear.

Box 1. COMPAS algorithm controversy

The COMPAS algorithm developed by the firm Equivant (formerly Northpointe) is commonly used by judges and forecasts which criminals are most likely to reoffend. It took a quantitative approach related to fairness, seeking to correctly predict recidivism for defendants and being as accurate as possible across all individuals. ProPublica found that while it did correctly predict recidivism for Black and white defendants at roughly the same rate, when it was wrong, it was wrong in different ways for Black and white people: Black arrestees who would not be rearrested in a 2-year horizon scored as high risk at twice the rate of white arrestees not subsequently arrested.[6] It also scored white people who were more likely than Black people to go on to commit a crime, as lower.[7] By doing this, the algorithm perpetuates a status quo, without incorporating how and why the policing system has and continues to be discriminatory against Black people.

Northepointe argued that the COMPAS algorithm was fair since the model reflected the same likelihood of recidivism across all groups. This type of reasoning reflects the rationale in government decision-making by treating all citizens according to the same rules. ProPublica argued that it is not fair in terms of treating likes alike (especially as race is a protected social group category). So it did not achieve the quantitative definition of fairness when it was wrong. This approach also violates other definitions of fairness - particularly from a social science and political philosophy perspective. But still, there is no clear "wrong" or "right". Questions that arise: *Is this the right framing of fairness to take in this context? Should a private sector actor be deciding what is fair in this public sector matter?*

III. Challenges

We trust ML systems expecting them to be "fair" and not discriminatory (especially not to discriminate against groups that are legally protected, such as by race and gender). But it's not that simple. Rarely are different definitions and notions of fairness considered at the start of developing an ML system. Even if different definitions or approaches are considered, there is not necessarily a "right" answer for a particular AI system. Also, there are various actors involved in the ML process (from dataset development through development of algorithms and use of AI systems) that might have different understandings and interpretations of fairness.

In the case of the COMPAS algorithm, the law provides some direction but still leaves room for interpration of what fairness means. Importantly, AI systems don't just mirror society but have the potential to replicate it over time and even amplify inequities that exist. Maintaining the status quo through AI systems - as the COMPAS algorithm is set up to do - has the potential to mask, perpetuate and amplify inequities. Again, there is no "right" answer and opinions differ on what's "fair" in that situation. Many advocates highlight the importance of centering on justice (see Box 2).

It can be easy and even tempting to check the box and say the model is 'fair' according to whatever definition or approach used. This can be problematic especially if it's done without clear explanation of why or how a particular approach was taken. Selecting a fairness definition / approach means making trade offs- and these trade offs need to be documented in order to understand what an AI system is designed to do and why — as well as allow for debate.

From a technical perspective, there are other challenges: adding more fairness constraints places restrictions on an algorithm resulting in lower accuracy.[8] Also, the opaqueness of ML models can make it challenging to ensure "fairness".[9]

Beyond construction of an ML system, fairness comes into play in terms of how the system is used. It can be considered unfair if users can't see, understand or appeal choices made by AI systems.[10]

Box 2. Moving towards promoting justice

Justice is centered around equity in every aspect of society. Society is a product of its history - with opportunities and resources informed and allocated along socially-constructed group lines (e.g., race, gender, class, sexual orientation, and ability). A justice approach recognizes how certain groups have been oppressed or marginalized, and seeks to address this to enhance freedom and possibility for all. When it comes to AI systems, a justice approach considers how certain groups are oppressed or marginalized in the particular context and explores how the AI system can advance equity, rather than perpetuate a status quo that may oppress or marginalize certain groups.

IV. Tools

Tools can help practitioners navigate the murky waters of fairness. They can provide guidance, help formalize processes, and empower individual employees. They also serve to document decisions so teams can clarify their position, as well as allow for debate.[11]

Technical / Quantitative tools

There are several AI fairness tools meant to help engineers and data scientists examine, report, and mitigate discrimination and bias in ML models. For example:

- IBM's AI Fairness 360 Toolkit: a Python toolkit focusing on technical solutions through fairness metrics and algorithms to help users examine, report, and mitigate discrimination and bias in ML models.
- Google's What-If Tool: a tool to explore a models' performance on a dataset, including examining several preset definitions of fairness constraints (e.g., equality of opportunity).[12] This tool is interesting as it allows users to explore different definitions of fairness.
- Microsoft's fairlean.py: a Python package that implements a variety of algorithms that seek to mitigate "unfairness" in supervised machine learning.
- Facebook is developing a "Fairness Flow" internal tool to identify bias in ML models.

Regardless of a focus on data or the broader AI system lifecycle, these tools tend to use a technical lens and focus on technical solutions. Technical solutions are important, but miss important fairness considerations. A tool employing purely technical solutions would not have captured the nuances behind the COMPAS algorithm's discrimination. A purely technical approach is insufficient to understand and mitigate biases. It perpetuates the misleading notion that ML systems can achieve "fairness" or be "un-biased".

Qualitative tools

Qualitative tools help delve into the nuances of fairness and prompt important discussion and reflection. They can enable teams to envision the AI system and its role in society, explore potential fairness-related harms and trade-offs, outline how bias could occur, and prepare plans to mitigate biases. They can also help track and monitor fairness-related harms that might come into play.[13]

We highlight two qualitative tools:

- Co-designed AI fairness checklist (2020): A group of Microsoft researchers and academic researchers engaged 48 individuals from 12 technology companies to co-design an AI fairness checklist. The checklist includes items to cover at the different stages of an AI system development and deployment lifecycle (i.e. envision, define, prototype, build, launch, and evolve). The checklist is meant to be customized.[14]
- Fairness Analytic (2019): This tool developed by Mulligan et al is designed to facilitate conversations about fairness during earlier stages of a project. It allows teams to explore concepts of fairness from various disciplines and think about what fairness could and should mean for a particular AI system. It helps teams understand what terms are being used, promote debate and develop a shared understanding.[15]

Like technical-oriented tools, they have limitations. For example, checklists can potentially be "gamed" – especially when there is a tendency at an organization to focus on technical solutions (explicitly or not).

While there are various tools that exist, it's important for users to be clear on which tools they are using and which gaps those tools address – and do not address. In many cases, several tools (spanning technical and non-technical solutions) are useful and needed.

V. Considerations

1. Identify fairness considerations and approaches up front, and ensure appropriate voices (i.e. experts in the relevant domain and across disciplines) are included and empowered in the conversation.
2. Instead of trying to make an ML system completely fair (or "de-biasing" it), the goal can be to detect and mitigate fairness-related harms as much as possible. Questions that should always be asked include: fair to whom? In what context?
3. There aren't always clear-cut answers, so document processes and considerations (including priorities and trade offs).
4. Use quantitative and qualitative approaches and tools to help facilitate these processes. Tools do not guarantee fairness! They are a good practice within the larger holistic approach to mitigating bias.
5. Fairness doesn't stop once an AI system is developed. Ensure users and stakeholders can see, understand and appeal choices made by AI systems.

This brief was written by Genevieve Smith with input and feedback from Nitin Kohli & Ishita Rustagi (2020). It is an accompanying resource of Mitigating Bias in Artificial Intelligence: An Equity Fluent Leadership Playbook of the Center for Equity, Gender & Leadership (EGAL) at Berkeley Haas.

ENDNOTES

1 Mulligan, D., Kroll, J., Kohli, N. & Wong, R. (2019). This thing called fairness: Disciplinary confusion realizing a value in technology. ACM Human-Computer Interaction, 3, 119. https://doi.org/10.1145/3359221.

2 For a more complete overview of "fairness" definitions and approaches across disciplines and how this relates to technology, see: Mulligan, D., Kroll, J., Kohli, N. & Wong, R. (2019). This thing called fairness: Disciplinary confusion realizing a value in technology. ACM Human-Computer Interaction, 3, 119. https://doi.org/10.1145/3359221.

3 Mulligan, et al (2019).

4 Mulligan, et al (2019).

5 Mulligan, et al (2019).

6 Larson, Jeff, Surya Mattu, Lauren Kirchner, and Julia Angwin. "How we analyzed the COMPAS recidivism algorithm." ProPublica (5 2016) 9, no. 1 (2016).

7 Spielkamp, M. (2017). Inspecting algorithms for bias. MIT Technology Review. Retrieved from https://www.technologyreview.com/2017/06/12/105804/inspecting-algorithms-for-bias/.

8 Mulligan, et al (2019).

9 IBM Research. An end-to-end machine learning pipeline that ensures fairness. https://arxiv.org/abs/1710.06876

10 Burrell, J., Mulligan, D. & Kluttz, D. (2018). Report from the first AFOG summer workshop. Algorithmic Fairness and Opacity Working Group (AFOG) at UC Berkeley. Retrieved from https://afog.berkeley.edu/files/2019/07/AFOG_workshop2018_report_all_web.pdf.

11 Cansu, C. (May 2, 2020). Personal interview.Burrell, J., Mulligan, D. & Kluttz, D. (2018). Report from the first AFOG summer workshop. Algorithmic Fairness and Opacity Working Group (AFOG) at UC Berkeley. Retrieved from https://afog.berkeley.edu/files/2019/07/AFOG_workshop2018_report_all_web.pdf.

12 Weinberger, D. Playing with fairness. Google. Retrieved from https://pair-code.github.io/what-if-tool/ai-fairness.html.

13 Madaio, M., Stark, L., Vaughan, J. W. & Wallach, H. (2020). Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. CHI 2020 paper. http://www.jennwv.com/papers/checklists.pdf.

14 Madaio, et al (2020).

15 Mulligan, et al (2019).