“20% Increase in fairness for Black applicants”: A Critical Examination of Fairness Measurements Offered by Startups

Corinna Hertweck[1,2], Maya Guido[1]

[1] *University of Zurich, Zurich, Switzerland*

[2] *Zurich University of Applied Sciences, Zurich, Switzerland*

Abstract

Companies using machine learning are increasingly obligated to integrate fairness considerations, often driven by regulatory imperatives and public discourse. This has given rise to a startup ecosystem focused on or at least integrating fairness measurement into their ML observability platforms. However, fairness is a complex concept and there are still many open questions in research. We therefore investigate how startups deal with this and present preliminary results of our ongoing analysis of the fairness startup landscape. In our analysis, we review publicly available material (such as websites) from these companies. We find two notable gaps: (1) the gap between fairness measurement in the algorithmic fairness literature and what startups actually implement and (2) the gap between the claims made by these startups and their actual practices. Based on our findings, we make recommendations for academia, policymakers, and industry stakeholders to advance the cause of fairness in machine learning collaboratively.

Keywords

fairness, observability, startups, fairness metrics, fairness criteria, demographic parity, statistical parity

1. Introduction

Through the increasing use of machine learning, there is also an increasing awareness of potential discrimination through automated decision-making systems. This has led to more regulation in this space (e.g., in the EU AI Act [1]) and thereby to more pressure on companies that are using machine learning. Consequently, ML observability platforms are starting to incorporate fairness metrics into their offerings. Some of these platforms even prioritize fairness as their primary concern. However, it is unclear if these platforms' claims match what they can actually offer – especially since we know that the field of algorithmic fairness still has a lot of open questions to answer on the research side. Inspired by [2], we want to evaluate these platforms' “claims and practices”. Our focus is specifically on startups that integrate some form of off-the-shelf fairness measurement into their platforms. We do not consider consulting companies that do not offer stand-alone platforms and instead provide services such as consultation or manual audits. For an overview of the AI audit ecosystem, we refer readers

EWAF'24: European Workshop on Algorithmic Fairness, July 01–03, 2024, Mainz, Germany

corinna.hertweck@zhaw.ch (C. Hertweck); maya.guido@uzh.ch (M. Guido)
https://hcorinna.github.io/ (C. Hertweck)
0000-0002-7639-2771 (C. Hertweck); 0000-0002-2770-1216 (M. Guido)

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

to [3]. We also do not consider open source platforms, which [4] has reviewed. Our goal is to provide an overview of the fairness measurement startup ecosystem and to discuss how these startups implement fairness measurement in practice. We aim to highlight the gaps between current implementations and existing research and suggest potential improvements in both research and implementation to guide algorithmic fairness in practice.

2. Methods

We collected relevant startups specializing in fairness evaluations from "The ethical AI database", Google search and Crunchbase, using a set of predefined keywords related to algorithmic fairness. We then filtered this list for startups that claim to offer fairness metrics. This resulted in a list of 21 startups, which we are currently investigating. Since their platforms are proprietary products, we were not able to easily access them to check what types of fairness measurements are implemented. We therefore rely on startups' publicly available material, such as their website, documentation, white papers and video material. We review this material to document how these startups implement fairness measurement and also take note of the claims that they are making about their products. The startups that we have analyzed so far are Arize [5], Etiq AI [6], FairPlay [7], Fiddler AI [8], Mona [9] and SolasAI [10].

3. Preliminary Results

3.1. Fairness Measurement

For Fiddler AI, Arize and Etiq AI, we were able to find a clear list of the implemented fairness criteria (see [11, 12, 13]). FairPlay uses one metric in all their reports, which we therefore assume is the only one that their platform measures although they mention two more metrics on their website's FAQ section [14]. For Mona and SolasAI, we could not find documentation that listed the implemented fairness metrics, so access to the platform would be required to evaluate this further. Note that these platforms also implement other metrics (e.g., label distribution) for evaluating different aspects. However, we focus specifically on fairness metrics and how users are guided to choose between them.

Focus on standard group fairness criteria Of the platforms with information on which concrete fairness criteria are implemented, all but one of the implemented criteria belong to the group fairness category. Only Etiq AI mentions individual fairness [13]. However, there is no explanation of how these are implemented or how the issue of defining similarity between individuals is addressed. All other implemented fairness metrics are group fairness metrics. This is a clear majority that resembles what we see in the open source landscape [4]. We assume that the reason for this is that group fairness is very easy to implement and requires no further input from users whereas individual fairness or causal definitions of fairness require domain-specific input from the user.

Implemented fairness criteria Let us now summarize which fairness criteria we know to be implemented.[1]

- **Statistical parity / demographic parity:** selection rate (probability of receiving a positive decision) equal across socio-demographic groups; implemented by all four startups
- **Equal opportunity:** true positive rate equal across socio-demographic groups; implemented by three startups (Fiddler AI, Arize, Etiq AI)
- **False positive rate parity:** false positive rate equal across socio-demographic groups; implemented by one startup (Arize)
- **Equalized odds:** both equal opportunity and false positive rate parity[2] fulfilled; implemented by one startup (Etiq AI)
- **Group benefit parity:** ratio of positive decisions to positive labels equal across socio-demographic groups; implemented by one startup (Fiddler AI)
- **Denial odds parity** ratio of negative decisions to positive decisions equal across socio-demographic groups. The ratio of two groups' denial odds is described as a fairness metric in FairPlay's FAQ section [14], but it is doubtful whether it is actually implemented.

The first four of these criteria are well-known group fairness criteria that are commonly found in the literature. However, they have also received criticism: One common theme is that these fairness criteria only look at statistics relating to the decision but not at the consequences of the decision [15, 16, 17, 18, 19]. However, what is relevant for fairness is how a decision affects decision subjects. This mismatch can mean that enforcing some fairness metrics could hurt marginalized groups as shown in [20, 19]. There has thus been a call for welfare-based fairness criteria, which the analyzed tools have not implemented yet.

Lack of guidance Choosing an appropriate fairness metric represents multiple value judgments about the situation at hand. This moral choice is difficult to make, but particularly hard if one is not familiar with fairness and justice discussions – which we would expect to be the case for practitioners using these platforms. We therefore sought documentation from all platforms that guide users in choosing fairness metrics. Along with the specification of the fairness metrics that are implemented, Fiddler AI, Arize, Etiq AI and FairPlay all provided more information on these metrics. However, in three cases (Fiddler AI, Etiq AI and FairPlay) this information is purely formal and descriptive. They simply describe the statistical metric in words instead of using a formula. What is provided is not actual guidance, but something that merely appears to be guidance at first. See, for example, Fiddler AI's "guidance" on two fairness criteria (the others are described similarly) in [11]:

- **Group benefit:** "If the two groups are treated equally, the group benefit should be the same."
- **Equal opportunity:** "If the two groups are treated equally, the TPR should be the same."

[1] Note that because we only have access to the documentation and white papers, but not the platforms themselves, there could be discrepancies that we cannot account for.

[2] Etiq AI actually uses equal opportunity and true negative rate parity, but by fulfilling true negative rate parity, one also fulfills false positive rate parity.

Wanting groups to be treated equally seems like a good goal, which according to Fiddler AI would mean having to fulfill both the group benefit and equal opportunity criterion – which Fiddler AI (incorrectly) claims to be “impossible”.[3] The given information is not only confusing to users but also not backed up by research.

In a blog post [23], Arize provides a decision tree through which users are supposed to find appropriate fairness criteria. This tree strongly resembles the one proposed by Aequitas [24].[4] With questions such as “Does your business problem require fairness to address disparate representation or disparate errors in your ML model?”, the tree would (similar to Aequitas’ tree, cmp. [4]) still be difficult to use for an uninitiated user of a fairness toolkit as they assume that a user already knows what fairness requires in their context.

With access limited to the platforms’ websites and documentation, it’s unclear if more guidance is available on the actual platforms. Given the unclear documentation, we do not expect this to be the case.

3.2. Critical View on Claims

In our analysis, we came across various claims about fairness measurement and bias mitigation capabilities of startups. Some startups give the impression that fairness is fully quantifiable with a definite metric to measure bias, even though a single fairness metric cannot capture the complexity of fairness. [25]. For bias mitigation, it is common to insinuate that mitigation techniques are a solution or fix for discrimination – a techno-solutionist message [26, 27]. One example that combines both is the following claim found on FairPlay’s website, advertising why customers should use FairPlay’s platform: “20% Increase in fairness for Black applicants” [28]. These kinds of claims carry the risk that third parties using these platforms build on the claims of the startups to ethics-wash their product.

4. Discussion

As we have seen, most implemented fairness metrics are standard group fairness metrics. While group fairness metrics have the advantage of being easy to implement, this also bears the danger that they are used without much reflection. This issue is worsened by the platform providers not offering any sort of moral guidance for choosing fairness metrics. Moreover, many startups make misleading claims about their fairness capabilities that promote a techno-solutionist view, reducing fairness to a single number. Although some startups have shown admirable intentions in practical fairness solutions, they are inherently driven by customer demand – which is in this case often a reaction to prevalent regulations. Therefore, achieving substantive fairness must be a collective responsibility that extends beyond these platforms and encompasses policymakers, researchers, the industry and society at large.

[3] Fiddler AI writes “An important point to make is that it’s impossible to optimize all the metrics at the same time. This is something to keep in mind when analyzing fairness metrics.” With this, Fiddler AI hints at the impossibility theorems [21, 22], which mathematically show the impossibility of fulfilling specific criteria at the same time under certain conditions. However, they only showed this impossibility for certain metrics and, for example, did not include group benefit.

[4] Although we note that the work of Aequitas is not cited by Arize.

References

[1] European Commission, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021) 206 final), 2021. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206, accessed on 2024-01-03.

[2] M. Raghavan, S. Barocas, J. Kleinberg, K. Levy, Mitigating bias in algorithmic hiring: Evaluating claims and practices, in: Proceedings of the 2020 conference on fairness, accountability, and transparency, 2020, pp. 469–481.

[3] S. Costanza-Chock, I. D. Raji, J. Buolamwini, Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 1571–1583.

[4] M. S. A. Lee, J. Singh, The landscape and gaps in open source fairness toolkits, in: Proceedings of the 2021 CHI conference on human factors in computing systems, 2021, pp. 1–13.

[5] Arize, The AI Observability & LLM Evaluation Platform, 2024. URL: https://arize.com/, accessed on 2024-03-28.

[6] Etiq AI, ML Testing For Everyone, 2024. URL: https://etiq.ai/, accessed on 2024-03-28.

[7] FairPlay, Fairness for People, Profits, and Progress, 2024. URL: https://fairplay.ai/, accessed on 2024-03-28.

[8] Fiddler AI, AI Observability, 2024. URL: https://www.fiddler.ai/, accessed on 2024-03-28.

[9] Mona, The Most Intelligent AI Monitoring Platform, 2023. URL: https://www.monalabs.io/, accessed on 2024-03-28.

[10] SolasAI, Reduce your algorithmic discrimination regulatory, legal and reputational risk, 2022. URL: https://www.solas.ai/, accessed on 2024-03-28.

[11] Fiddler AI, Fairness, 2023. URL: https://docs.fiddler.ai/docs/fairness, accessed on 2023-11-13.

[12] Arize, Bias Tracing (Fairness), 2023. URL: https://docs.arize.com/arize/tracing-and-troubleshooting/11.-bias-tracing-fairness, accessed on 2023-11-25.

[13] Etiq AI, Bias, 2023. URL: https://docs.etiq.ai/scan-types/bias, accessed on 2023-12-01.

[14] FairPlay, Frequently Asked Questions, 2024. URL: https://fairplay.ai/faq/, accessed on 2024-02-26.

[15] J. Finocchiaro, R. Maio, F. Monachou, G. K. Patro, M. Raghavan, A.-A. Stoica, S. Tsirtsis, Bridging machine learning and mechanism design towards algorithmic fairness, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 489–503.

[16] R. Binns, Fairness in machine learning: Lessons from political philosophy, in: S. A. Friedler, C. Wilson (Eds.), Proceedings of the 1st Conference on Fairness, Accountability and Transparency, volume 81 of *Proceedings of Machine Learning Research*, PMLR, New York, NY, USA, 2018, pp. 149–159. URL: http://proceedings.mlr.press/v81/binns18a.html.

[17] C. Hertweck, C. Heitz, M. Loi, On the moral justification of statistical parity, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 747–757. URL: https://doi.org/10.1145/3442188.3445936. doi:10.1145/3442188.3445936.

[18] H. Weerts, L. Royakkers, M. Pechenizkiy, Does the End Justify the Means? On the Moral Justification of Fairness-Aware Machine Learning, arXiv preprint arXiv:2202.08536 (2022).

[19] M. Jorgensen, H. Richert, E. Black, N. Criado, J. Such, Not so fair: The impact of presumably fair machine learning models, in: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, 2023, pp. 297–311.

[20] L. Hu, Y. Chen, Fair classification and social welfare, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 535–545.

[21] J. Kleinberg, S. Mullainathan, M. Raghavan, Inherent trade-offs in the fair determination of risk scores, arXiv preprint arXiv:1609.05807 (2016).

[22] A. Chouldechova, Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, Big data 5 (2017) 153–163.

[23] S.-A. DeLucia, Evaluating Model Fairness, 2023. URL: https://arize.com/blog/evaluating-model-fairness/.

[24] Center for Data Science and Public Policy, University of Chicago, Aequitas, 2018. URL: http://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/.

[25] A. Z. Jacobs, H. Wallach, Measurement and fairness, in: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021, pp. 375–385.

[26] E. Morozov, To save everything, click here: The folly of technological solutionism, PublicAffairs, 2013.

[27] R. Abebe, S. Barocas, J. Kleinberg, K. Levy, M. Raghavan, D. G. Robinson, Roles for computing in social change, in: Proceedings of the 2020 conference on fairness, accountability, and transparency, 2020, pp. 252–260.

[28] FairPlay, Increase Fairness, Boost Profits, 2024. URL: https://fairplay.ai/for-banks/, accessed on 2024-03-28.