The Need for Top-Level Performance Indicators to Support Fairness in AI

Michael A. Madaio
Carnegie Mellon University
Pittsburgh, PA, USA
mmadaio@cs.cmu.edu

Luke Stark
Microsoft Research
Montreal, Canada
luke.stark@microsoft.com

Jennifer Wortman Vaughan
Microsoft Research
New York, NY, USA
jenn@microsoft.com

Hanna Wallach
Microsoft Research
New York, NY, USA
wallach@microsoft.com

CHI'20,, April 25–30, 2020, Honolulu, HI, USA
ACM 978-1-4503-6819-3/20/04.
https://doi.org/10.1145/3334480.XXXXXXX

Abstract

Following the recent publication of dozens of AI ethics statements and responsible AI principles, technologists have begun to operationalize concepts such as fairness into metrics, toolkits, and checklists in order to impact AI product development. However, although AI practitioners may wish to use such methods to develop fairer AI products and services, there are organizational incentives that inhibit individuals from advocating for and addressing fairness issues in practice. In this workshop paper, we present new findings from an AI fairness checklist co-design research project [6] that suggest future directions and open questions for developing top-level performance indicators to support AI fairness efforts, focusing specifically on the challenges of designing such indicators so that they are both effective and legible to organizations. We intend for this paper to spark a discussion around aligning organizational incentives to support the development of fairer AI products and services.

Author Keywords

AI; ML; ethics; fairness; co-design; checklists

CCS Concepts

•**Human-centered computing** → **Collaborative and social computing;** •**Social and professional topics** → *Codes of ethics;* •**Computing methodologies** → *Machine learning;*

Introduction

Artificial intelligence (AI) technologies are increasingly ubiquitous, embedded in products and services throughout many different sectors. Although AI has enormous potential for good, it can also amplify and reify existing societal injustices [7]. To mitigate such harms, many public- and private-sector organizations have published AI ethics statements and responsible AI principles intended to impact AI product development (see Jobin et al. [3] for a review). To operationalize concepts such as fairness, for example, technologists have begun to create metrics and toolkits (e.g., [8]), as well as checklists (e.g., [5]). Indeed, checklists are increasingly touted as a potential solution for AI practitioners who wish to avoid the development of unfair products and services. However, individuals belonging to larger product teams may be unable to use these checklists to impact product development, despite their best intentions [1, 2].

To better understand the role of checklists in developing fairer AI products and services, we co-designed (cf. [10]) an AI fairness checklist with 48 AI practitioners, and identified desiderata and concerns for AI fairness checklists more broadly [6]. We found that practitioners felt that there are organizational incentives that inhibit individuals from advocating for and addressing fairness issues, but that checklists could provide infrastructure to support such efforts. In particular, despite individuals' willingness to prioritize fairness, organizations' incentives to ship quickly were at odds with the time required to carefully consider fairness issues.

In this workshop paper, we build on this research by describing additional findings that suggest future directions and open questions for developing top-level performance indicators to support AI fairness efforts—questions that we hope to discuss with other workshop participants.

Findings and Insights

Practitioners want KPIs that reflect AI fairness efforts

We found that many AI practitioners want some definition of fairness to be one of their organization's top-level performance indicators, variously called key performance indicators (KPIs), objectives and key results (OKRs), or scorecards [4]. Many organizations use top-level performance indicators in their AI product development processes to evaluate product readiness and inform employee promotion [9]. Practitioners told us that when top-level goals were clearly visible, along with target thresholds or criteria for achieving them, they were better able to advocate for resources or additional development time. Furthermore, practitioners described cases where senior leadership would either publicly praise team members for meeting target thresholds or, alternatively, delay promotions for not meeting them.

We also found that although many organizations have top-level performance indicators for accuracy, revenue, speed of shipping, security, privacy, etc., they largely lack comparable indicators for fairness. However, practitioners told us that they *wanted* top-level performance indicators for fairness—either for their products and services or for their processes for addressing fairness issues. Indeed, other product development performance indicators include evaluations of products and services, as well as product development processes [4, 9]. We found that some teams are already working to create top-level performance indicators for concepts that are related to fairness, such as inclusiveness or demographic representativeness. Practitioners described how the process of creating these indicators prompted them to articulate what it meant for their product or service to be inclusive or representative in such a way that they could then measure and track their performance over time.

Challenges in developing AI fairness KPIs

Some practitioners were uncertain about whether fairness could be measured, while others wanted to avoid binary thinking about fairness (i.e., the product or service either is or isn't fair), questioning whether it was possible to develop target thresholds for fairness (cf. [4]). Several practitioners used an example of a three-color (red, yellow, green) scheme for top-level performance indicators for security to demonstrate that, even in cases where a product or service may not be completely secure, the team had conducted a high-fidelity assessment of potential security threats. Practitioners felt that an analogous scheme would allow them to communicate their progress in addressing fairness issues to senior leadership, advocate for resources or additional development time, and support employee promotion.

Practitioners reported that although having top-level performance indicators for fairness would be useful, they were concerned about the difficulty of developing such indicators. Although some practitioners drew analogies between fairness and security, they also described how their organizations took many years to develop top-level performance indicators for security. Practitioners were also worried about integrating or reconciling online, quantitative indicators with offline, qualitative evaluations of their products and services. Some cited an example of user engagement and web search, describing how qualitative evaluations revealed that users were unhappy about seeing adult content, despite high click-through rates. Practitioners also described challenges in aligning the cadence of their online and offline evaluations. Although offline, qualitative evaluations are used prior to deployment, many teams primarily use online, quantitative indicators after deployment—rarely (if ever) continuing to conduct qualitative evaluations with users.

Discussion

These findings suggest that future work is needed to understand how organizations can incentivize practitioners to advocate for and address fairness issues. This may involve developing top-level performance indicators for fairness, thereby enabling teams to measure and track their performance over time. However, it is not clear how fairness should be measured. Moreover, there are major risks in reducing complex, societal concepts like fairness to indicators. Indicators may fail to capture some types of fairness issues, and can be gamed. Finally, top-level performance indicators may not be appropriate for scenarios where the fairest decision is to not develop a product or service at all.

We hope this paper will foster a discussion with other workshop participants, through which we can collectively explore future directions for aligning organizational incentives to support the development of fairer AI products and services.

REFERENCES

[1] Colin M Gray and Shruthi Sai Chivukula. 2019. Ethical Mediation in UX Practice. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 178. DOI: http://dx.doi.org/10.1145/3290605.3300408

[2] Daniel Greene, Anna Lauren Hoffmann, and Luke Stark. 2019. Better, Nicer, Clearer, Fairer. In *Proceedings of the 52nd Hawaii International Conference on System Sciences (HICSS)*, T X Bui and R H Sprague (Eds.). 2122–2131.

[3] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* (Sept. 2019), 1–11.

[4] Lj Kazi, Biljana Radulovic, and Zoltan Kazi. 2012. Performance indicators in software project monitoring:

Balanced scorecard approach. In *2012 IEEE 10th Jubilee International Symposium on Intelligent Systems and Informatics*. IEEE, 19–25.

[5] Mike Loukides, Hilary Mason, and DJ Patil. 2018. Of Oaths and Checklists. (2018). https://www.oreilly.com/ideas/of-oaths-and-checklists

[6] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.

[7] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society* 3, 2 (Sept. 2016), 205395171667967–21.

[8] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, and Rayid Ghani. 2018. Aequitas: A Bias and Fairness Audit Toolkit. 2018 (2018). http://arxiv.org/abs/1811.05577

[9] Miroslaw Staron, Wilhelm Meding, and Klas Palm. 2012. Release readiness indicator for mature agile and lean software development projects. In *International Conference on Agile Software Development*. Springer, 93–107.

[10] Daisy Yoo, Alina Huldtgren, Jill Palzkill Woelfer, David G Hendry, and Batya Friedman. 2013. A value sensitive action-reflection model: evolving a co-design space with stakeholder and designer prompts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 419–428. DOI: http://dx.doi.org/10.1145/2470654.2470715