Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products

Inioluwa Deborah Raji
University of Toronto
27 King's College Cir
Toronto, Ontario, Canada, M5S 3H7
deborah.raji@mail.utoronto.com

Joy Buolamwini
Massachusetts Institute of Technology
77 Massachusetts Ave
Cambridge, Massachusetts, 02139
joyab@mit.edu

Abstract

Although algorithmic auditing has emerged as a key strategy to expose systematic biases embedded in software platforms, we struggle to understand the real-world impact of these audits, as scholarship on the impact of algorithmic audits on increasing algorithmic fairness and transparency in commercial systems is nascent. To analyze the impact of publicly naming and disclosing performance results of biased AI systems, we investigate the commercial impact of Gender Shades, the first algorithmic audit of gender and skin type performance disparities in commercial facial analysis models. This paper 1) outlines the audit design and structured disclosure procedure used in the Gender Shades study, 2) presents new performance metrics from targeted companies IBM, Microsoft and Megvii (Face++) on the Pilot Parliaments Benchmark (PPB) as of August 2018, 3) provides performance results on PPB by non-target companies Amazon and Kairos and, 4) explores differences in company responses as shared through corporate communications that contextualize differences in performance on PPB. Within 7 months of the original audit, we find that all three targets released new API versions. All targets reduced accuracy disparities between males and females and darker and lighter-skinned subgroups, with the most significant update occurring for the darker-skinned female subgroup, that underwent a 17.7% - 30.4% reduction in error between audit periods. Minimizing these disparities led to a 5.72% to 8.3% reduction in overall error on the Pilot Parliaments Benchmark (PPB) for target corporation APIs. The overall performance of non-targets Amazon and Kairos lags significantly behind that of the targets, with error rates of 8.66% and 6.60% overall, and error rates of 31.37% and 22.50% for the darker female subgroup, respectively.

Introduction

An algorithmic audit involves the collection and analysis of outcomes from a fixed algorithm or defined model within a system. Through the stimulation of a mock user population, these audits can uncover problematic patterns in models of interest. Targeted public algorithmic audits provide one mechanism to incentivize corporations to address the algorithmic bias present in data-centric technologies that continue to play an integral role in daily life, from governing access to information and economic opportunities to influencing personal freedoms (Julia Angwin and Kirchner 2016; Jakub Mikians 2012; Aniko Hannak and Wilson 2017; Edelman and Luca 2014).

However, researchers who engage in algorithmic audits risk breaching company Terms of Service, the Computer Fraud and Abuse Act (CFAA) or ACM ethical practices as well as face uncertainty around hostile corporate reactions. Given these risks, much algorithmic audit work has focused on goals to gauge user awareness of algorithmic bias (Eslami et al. 2017; Kevin Hamilton and Sandvig 2015) or evaluate the impact of bias on user behaviour and outcomes (Gary Soeller and Wilson 2016; Juhi Kulshrestha 2017; Edelman and Luca 2014), instead of directly challenging companies to change commercial systems. Research on the real-world impact of an algorithmic audit is thus needed to inform strategies on how to engage corporations productively in addressing algorithmic bias. The Buolamwini & Gebru Gender Shades study (Buolamwini and Gebru 2018), which investigated the accuracy of commercial gender classification services, provides an apt case study to explore audit design and disclosure practices that engage companies in making concrete process and model improvements to address classification bias in their offerings.

Related Work

Corporations and Algorithmic Accountability

As more artificial intelligence (AI) services become mainstream and harmful societal impacts become increasingly apparent (Julia Angwin and Kirchner 2016; Jakub Mikians 2012), there is a growing need to hold AI providers accountable. However, outside of the capitalist motivations of economic benefit, employee satisfaction, competitive advantage, social pressure, and recent legal developments like the EU General Data Protection Regulation, corporations still have little incentive to disclose details about their systems (Diakopoulos 2016; Burrell 2016; Sandra Wachter and Russell 2018). Thus external pressure remains a necessary approach to increase transparency and address harmful model bias.

If we take the framing of algorithmic bias as a software defect or bug that poses a threat to user dignity or access to opportunity (Tramèr et al. 2015), then we can anticipate parallel challenges to that faced in the field of information security, where practitioners regularly address and communicate threats to user safety. The National Computer Emergency Readiness Team (CERT) promotes a strict procedure named "Coordinated Vulnerability Disclosures (CVD)" to inform corporations of externally identified cyber security threats in a way that is non-antagonistic, respectful of general public awareness and careful to guard against corporate inaction (Allen D. Householder and King 2017). CVDs outline the urgent steps of discovery, reporting, validation and triage, remediation and then subsequent public awareness campaigns and vendor re-deployment of a system identified internally or externally to pose a serious cyber threat. A similar "Coordinated Bias Disclosure" procedure could support action-driven corporate disclosure practices to address algorithmic bias as well.

Black Box Algorithmic Audits

For commercial systems, the audit itself is characterized as a "black box audit", where the direct or indirect influence of input features on classifier accuracy or outcomes is inferred through the evaluation of a curated benchmark (Philip Adler and Venkatasubramanian 2018; Riccardo Guidotti and Giannotti 2018). Benchmark test sets like FERET (P.J. Phillips and Rauss 2000) and the Facial Recognition Vendor Test (FRVT) from the National Institute of Standards and Technology (NIST) (Mei Ngan and Grother 2015) are of particular interest, as examples specific to establishing policy and legal restrictions around mitigating bias in facial recognition technologies.

In several implemented audit studies, vendor names are kept anonymous (Brendan F. Klare and Jain 2012) or the scope is scaled down to a single named target (Snow 2018; Le Chen and Wilson 2015; Juhi Kulshrestha 2017). The former fails to harness public pressure and the latter fails to capture the competitive dynamics of a multi-target audit - thus reducing the impetus for corporate reactions to those studies.

Gender Shades

The Gender Shades study differs from these previous cases as an external and multi-target black box audit of commercial machine learning Application Program Interfaces (APIs), scoped to evaluating the facial analysis task of binary gender classification (Buolamwini and Gebru 2018). The contribution of the work is two-fold, serving to introduce the gender and skin type balanced Pilot Parliaments Benchmark (PPB) and also execute an intersectional demographic and phenotypic evaluation of face-based gender classification in commercial APIs. The original authors consider each API's model performance given the test image attributes of gender, reduced to the binary categories of male or female, as well as binary Fitzpatrick score, a numerical classification schema for human skin type evaluated by a dermatologist, and grouped into classes of lighter and darker skin types. The audit then evaluates model performance across these unitary subgroups (i.e. female or darker) in addition to intersectional subgroups (i.e darker female), revealing large disparities in subgroup classification accuracy particularly across intersectional groups like darker female, darker male, lighter female and lighter male.

Analysis of Gender Shades Audit

Gender Shades Coordinated Bias Disclosure

In the Gender Shades study, the audit entity is independent of target corporations or its competitors and serves as a neutral 'third-party' auditor, similar to the expectation for corporate accounting auditing committees (Allen D. Householder and King 2017).

This neutrality enabled auditors to approach audited corporations systematically, following a procedure sequentially outlined below that closely mirrors key recommendations for coordinated vulnerability disclosures (CVDs) in information security (Allen D. Householder and King 2017).

1. **Documented Vulnerability Discovery** - A stated objective of the Gender Shades study is to document audit outcomes from May 2017 to expose performance vulnerabilities in commercial facial recognition products (Buolamwini and Gebru 2018).
2. **Defined Corporate Response Period with Limited Anonymomized Release to Audit Targets** - The Gender Shades paper (without explicit company references) was sent to Microsoft, IBM and Face++ on December 19th 2017 (Buolamwini 2017), giving companies prior notice to react before a communicated public release date, while maintaining the strict privacy of other involved stakeholders.
3. **Unrestricted Public Release Including Named Audit Targets** - On February 9th, 2018, "Facial Recognition Is Accurate, if You're a White Guy" , an article by Steve Lohr in the technology section of The New York Times is among the first public mentions of the study (Buolamwini 2017; 2018), and links to the published version of the study in Proceedings of Machine Learning Research, with explicit company references. This follows CVD procedures around alerting the public of corporate vulnerabilities with explicit culprit references, following a particular grace period in which companies are allowed to react before wider release. The Gender Shades public launch, accompanied by a video, summary visualizations and a website further prompts public, academic and corporate audiences - technical and non-technical alike - to be exposed to the issue and respond. Finally, the paper was presented on February 24th 2018 with explicit company references at the FAT* conference to an audience of academics, industry stakeholders and policymakers (Buolamwini 2017).
4. **Joint Public Release of Communications and Updates from Corporate Response Period** - Even if the issue is resolved, CVD outlines a process to still advance with the public release while also reporting corporate communications and updates from the response period. In the case

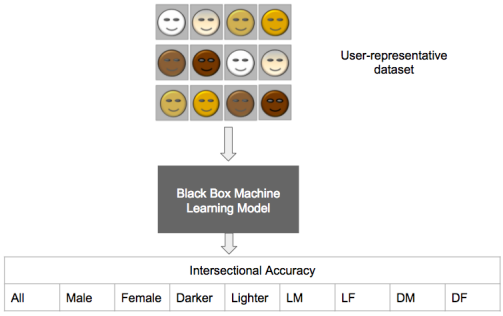

Figure 1: Gender Shades audit process overview (Buolamwini and Gebru 2018).

of Gender Shades, the co-author presented and linked to IBMs updated API results at the time of the public release of the initial study (Buolamwini 2017).

Gender Shades Audit Design

The Gender Shades paper contributed to the computer vision community the Pilot Parliaments Benchmark (PPB). PPB is an example of what we call a "user-representative" test set, meaning the benchmark does not have proportional demographic distribution of the intended user population but representative inclusion of the diversity of that group. With equal representation of each distinct subgroup of the user population regardless of the percentage at which that population is present in the sample of users, we can thus evaluate for equitable model performance across subgroups. Similar to the proposed error profiling under the Unwarranted Associations framework (Tramèr et al. 2015), algorithmic unfairness is evaluated by comparing classification accuracy across identified subgroups in the user-representative test set (see Figure 1).

Another key element of the audit design is that the audit targets are commercial machine learning Application Program Interfaces (APIs). The auditors thus mirror the behaviour of a single developer user for a commercial API platform that supports the creation of applications for the end user. Therefore, the actor being puppeted (the developer) has control of the application being used by the end-user and is at risk of propagating bias unto the end-users of their subsequent products. This is analogous to the "sock puppet" algorithmic audit model (Christian Sandvig and Langbort 2014) in that we pose as a puppet user of the API platform and interact with the API in the way a developer would. However, as this is a puppet that influences the end user experience, we label them "carrier puppets", acknowledging that, rather than evaluating a final state, we are auditing the bias detected in an intermediary step that can carry bias forward towards end users (see Figure 2).

Methodology

The design of this study is closely modeled after that of Gender Shades. Target corporations were selected from the original study, which cites considerations such as the platform

Figure 2: "Carrier puppet" audit framework overview.

market share, the availability of desired API functions and overall market influence as driving factors in the decision to select Microsoft, IBM and Face++ (Buolamwini and Gebru 2018). Non-target corporation Kairos was selected because of the company's public engagement with the Gender Shades study specifically and the topic of intersectional accuracy in general after the audit release (Brackeen 2018a; 2018b). Non-target corporation Amazon was selected following the revelation of the active use and promotion of its facial recognition technology in law enforcement (Cagle and Ozer 2018).

The main factor in analysis, the follow up audit, closely follows the procedure for the initial Gender Shades study. We calculated the subgroup classification error, as defined below, to evaluate disparities in model performance across identified subgroups, enabling direct comparison between follow-up results and initial audit results.

Subgroup Classification Error. Given data set $D = (X, Y, C)$, a given sample input d_i from D belongs to a subgroup S, which is a subset of D defined by the protected attributes X. We define black box classifier $g : X, Y \mapsto c$, which returns a prediction c from the attributes x_i and y_i of a given sample input d_i from D. If a prediction is not produced (i.e. face not detected), we omit the result from our calculations.

We thus define $err(S)$ be the error of the classifier g for members d_i of subgroup S to be as follows:

$$1 - P(g(x_i, y_i) = C_i | d_i \in S)$$

To contextualize audit results and examine language themes used post-audit, we considered written communications for all mentioned corporations. This includes exclusively corporate blog posts and official press releases, with the exception of media published corporate statements, such as an op-ed by the Kairos CEO published in TechCrunch (Brackeen 2018b). Any past and present website copy or Software Developer Kit documentation was also considered when determining alignment with identified themes, though this did not factor greatly into the results.

Performance Results

With the results of the follow up audit and original Gender Shades outcomes, we first analyze the differences between the performance of the targeted platforms in the original study and compare it to current target API performance. Next, we look at non-target corporations Kairos and Amazon, which were not included in the Gender Shades study and compare their current performance to that of targeted platforms.

Table 1: Overall Error on Pilot Parliaments Benchmark, August 2018 (%)

Company	All	Females	Males	Darker	Lighter	DF	DM	LF	LM
Target Corporations									
Face ++	1.6	2.5	0.9	2.6	0.7	4.1	1.3	1.0	0.5
MSFT	0.48	0.90	0.15	0.89	0.15	1.52	0.33	0.34	0.00
IBM	4.41	9.36	0.43	8.16	1.17	16.97	0.63	2.37	0.26
Non-Target Corporations									
Amazon	8.66	18.73	0.57	15.11	3.08	31.37	1.26	7.12	0.00
Kairos	6.60	14.10	0.60	11.10	2.80	22.50	1.30	6.40	0.00

Table 2: Overall Error Difference Between August 2018 and May 2017 PPB Audit (%)

Company	All	Females	Males	Darker	Lighter	DF	DM	LF	LM
Face ++	-8.3	-18.7	0.2	-13.9	-3.9	-30.4	0.6	-8.5	-0.3
MSFT	-5.72	-9.70	-2.45	-12.01	-0.45	-19.28	-5.67	-1.06	0.00
IBM	-7.69	-10.74	-5.17	-14.24	-1.93	-17.73	-11.37	-4.43	-0.04

The reported follow up audit was done on August 21, 2018, for all corporations in both cases. Summary Table 1 and Table 2 show percent error on misclassified faces of all processed faces, with undetected faces being discounted. Calculation details are outlined in the definition for Subgroup Classification Error and error differences are calculated by taking August 2018 error (%) and subtracting May 2017 error (%). DF is defined as darker female subgroup, DM is darker male, LM is lighter male and LF is lighter female.

Target Corporation Key Findings

The target corporations from the Gender Shades study all released new API versions, with a reduction in overall error on the Pilot Parliamentary Benchmark by 5.7%, 8.3% and 7.7% respectively for Microsoft, Face++ and IBM. Face++ took the most days to release their new API in 190 days (Face++ 2018), while IBM was the first to release a new API version in 66 days (Puri 2018), with Microsoft updating their product the day before Face++, in 189 days (Roach 2018). All targeted classifiers in post-audit releases have their largest error rate for the darker females subgroup and the lowest error rate for the lighter males subgroup. This is consistent with 2017 audit trends, barring Face++ which had the lowest error rate for darker males in May 2017.

The following is a summary of substantial performance changes across demographic and phenotypic classes, as well as their intersections, after API updates :

- Greater reduction in error for female faces (9.7% - 18.7% reduction in subgroup error) than male faces (0.2% - 5.17% reduction in error) .
- Greater reduction in error for darker faces (12.01% - 14.24% reduction in error) than for lighter faces (0.45% - 3.9% reduction in error).
- Lighter males are the least improved subgroup (0% - 0.3% reduction in error)
- Darker females are the most improved subgroup (17.7% - 30.4% reduction in error)
- If we define the error gap to be the error difference between worst and best performing subgroups for a given API product, IBM reduced the error gap from 34.4% to 16.71% from May 2017 to August 2018. In the same period, Microsoft closed a 20.8% error gap to a 1.52% error difference, and Face++ went from a 33.7% error gap to a 3.6% error gap.

Non-Target Corporation Key Findings

Non-target corporations Kairos and Amazon have overall error rates of 6.60% and 8.66% respectively. These are the worst current performances of the companies analyzed in the follow up audit. Nonetheless, when comparing to the previous May 2017 performance of target corporations, the Kairos and Amazon error rates are lower than the former error rates of IBM (12.1%) and Face++ (9.9%), and only slightly higher than Microsofts performance (6.2%) from the initial study. Below is a summary of key findings for non-target corporations:

- Kairos and Amazon perform better on male faces than female faces, a trend also observed in (Buolamwini and Gebru 2018; Mei Ngan and Grother 2015).
- Kairos and Amazon perform better on lighter faces than darker faces, a trend also observed in (Buolamwini and Gebru 2018; Jonathon Phillips and OToole 2011).
- Kairos (22.5% error) and Amazon (31.4% error) have the current worst performance for the darker female subgroup.
- Kairos and Amazon (both 0.0% error) have the current best performance for the lighter male subgroup.
- Kairos has an error gap of 22.5% between highest and lowest accuracy intersectional subgroups, while Amazon has an error gap of 31.37%.

Discussion

Given a clear understanding of the Gender Shades study procedure and follow up audit metrics, we are able to reflect on corporate reactions in the context of these results, and evaluate the progress made by this audit in influencing corporate action to address concerns around classification bias.

Reduced Performance Disparities Between Intersectional User Subgroups

Building on Crenshaw's 1989 research on the limitations of only considering single axis protected groups in anti-discrimination legislation (Crenshaw 1989), a major focus of the Gender Shades study is championing the relevance of intersectional analysis in the domain of human-centered AI systems. IBM and Microsoft, who both explicitly reference Gender Shades in product update releases, claim intersectional model improvements on their gender classifier (Puri 2018; Roach 2018). These claims are substantiated by the results of the August 2018 follow up audit, which reveals universal improvement across intersectional subgroups for all targeted corporations. We also see that the updated releases of target corporations mostly impact the least accurate subgroup (in this case, darker females). Although post-audit performance for this subgroup is still the worst relative to other intersectional subgroups across all platforms, the gap between this subgroup and the best performing subgroup - consistently lighter males - reduces significantly after corporate API update releases.

Additionally, with a 5.72% to 8.3% reduction in overall error on the Pilot Parliaments Benchmark (PPB) for target corporations, we demonstrate that minimizing subgroup performance disparities does not jeopardize overall model performance but rather improves it, highlighting the alignment of fairness objectives to the commercial incentive of improved qualitative and quantitative accuracy. This key result highlights an important critique of the current model evaluation practice of using a subset of the model training data for testing, by demonstrating the functional value in testing the model on a separately defined "user representative" test set.

Corporate Prioritization

Although the original study (Buolamwini and Gebru 2018) expresses the concern that potential physical limitations of the image quality and illumination of darker skinned subjects may be contributing to the higher error rate for that group, we can see through the 2018 performance results that these challenges can be overcome. Within 7 months, all targeted corporations were able to significantly reduce error gaps in the intersectional performance of their commercial APIs, revealing that if prioritized, the disparities in performance between intersectional subgroups can be addressed and minimized in a reasonable amount of time.

Several factors may have contributed to this increased prioritization. The unbiased involvement of multiple companies may have served to put capitalist pressure on each corporation to address model limitations as not to be left behind or called out. Similarly, increased corporate and consumer awareness on the issue of algorithmic discrimination and classification bias in particular may have incited urgency in pursuing a product update. This builds on literature promoting fairness through user awareness and education (Kevin Hamilton and Eslami 2014) - aware corporations can also drastically alter the processes needed to reduce bias in algorithmic systems.

Emphasis on Data-driven Solutions

These particular API updates appear to be data-driven. IBM publishes the statement "AI systems are only as effective as the data they're trained on" and both Microsoft and Kairos publish similar statements (Puri 2018; Roach 2018; Brackeen 2018a) , implying heavily the claim that data collection and diversification efforts play an important role in improving model performance across intersectional subgroups. This aligns with existing research (Irene Chen and Sontag 2018) advocating for increasing the diversity of data as a primary approach to improve fairness outcomes without compromising on overall accuracy. Nevertheless, the influence of algorithmic changes, training methodology or specific details about the exact composition of new training datasets remain unclear in this commercial context - thus underscoring the importance of work on open source models and datasets that can be more thoroughly investigated.

Non-technical Advancements

In addition to technical updates, we observe organizational and systemic changes within target corporations following the Gender Shades study. IBM published its "Principles for Trust and Transparency" on May 30th 2018 (IBM 2018), while Microsoft created an "AI and Ethics in Engineering and Research (AETHER) Committee, investing in strategies and tools for detecting and addressing bias in AI systems" on March 29th, 2018 (Smith 2018). Both companies also cite their involvement in Partnership for AI, an AI technology industry consortium, as a means of future ongoing support and corporate accountability (Puri 2018; Smith 2018).

Implicitly identifying the role of the API as a "carrier" of bias to end users, all companies also mention the importance of developer user accountability, with Microsoft and IBM speaking specifically to user engagement strategies and educational material on fairness considerations for their developer or enterprise clients (Puri 2018; Roach 2018).

Only Microsoft strongly mentions the solution of Diversity & Inclusion considerations in hiring as an avenue to address issues[38]. The founder of Kairos specifically claims his minority identity as personal motivation for participation in this issue, stating "I have a personal connection to the technology,...This resonates with me very personally as a minority founder in the face recognition space" (Brackeen 2018a; 2018b). A cultural shift in the facial recognition industry could thus attract and retain those paying increased attention to the issue due to personal resonance.

Differences between Target and Non Target Companies

Although prior performance for non-target companies is unknown, and no conclusions can be made about the rate of

product improvements, Kairos and Amazon both perform more closely to the target corporations' pre-audit performance than their post-audit performance.

Amazon, a large company with an employee count and revenue comparable to the target corporations IBM and Microsoft, seems optimistic about the use of facial recognition technology despite current limitations. In a response to a targeted ACLU audit of their facial recognition API (Snow 2018), they state explicitly, "Our quality of life would be much worse today if we outlawed new technology because some people could choose to abuse the technology". On the other hand, Kairos, a small privately held company not explicitly referenced in the Gender Shades paper and subsequent press discussions, released a public response to the initial Gender Shades study and seemed engaged in taking the threat of algorithmic bias quite seriously (Buolamwini and Gebru 2018).

Despite the varying corporate stances and levels of public engagement,the targeted audit in Gender Shades was much more effective in reducing disparities in target products than non-targeted systems.

Regulatory Communications

We additionally encounter scenarios where civil society organizations and government entities not explicitly referenced in the Gender Shades paper and subsequent press discussions publicly reference the results of the audit in letters, publications and calls to action. For instance, the Gender Shades study is cited in an ACLU letter to Amazon from shareholders requesting its retreat from selling and advertising facial recognition technology for law enforcement clients (Arjuna Capital 2018). Similar calls for action to Axon AI by several civil rights groups, as well as letters from Senator Kamala D. Harris to the EEOC, FBI and FTC regarding the use of facial recognition in law enforcement also directly reference the work (Coldewey 2017). Kairos, IBM and Microsoft all agree facial analysis technology should be restricted in certain contexts and demonstrate support for government regulation of facial recognition technology (IBM 2018; Smith 2018; Brackeen 2018b). In fact, Microsoft goes so far as to explicitly support public regulation (Smith 2018). Thus in addition to corporate reactions, future work might explore the engagement of government entities and other stakeholders beyond corporate entities in response to public algorithmic audits.

Design Considerations

Several design considerations also present opportunities for further investigation. As mentioned in Gender Shades, a consideration of confidence scores on these models is necessary to get a complete view on defining real-world performance (Buolamwini and Gebru 2018). For instance, IBM's self-reported performance on a replicated version of the Gender Shades audit claims a 3.46% overall error rate on their lowest accuracy group of darker females (Puri 2018) - this result varies greatly from the 16.97% error rate we observe in our follow up audit. Upon further inspection, we see that they only include results above a 99% confidence threshold whereas Gender Shades takes the binary label with the higher confidence score to be the predicted gender. These examples demonstrate the need to consider variations in results due to prediction confidence thresholding in future audit designs.

Another consideration is that the Gender Shades publication includes all the required information to replicate the benchmark and test models on PPB images (Buolamwini and Gebru 2018). It is possible that well performing models do not truly perform well on other diverse datasets outside of PPB and have been overfit to optimize their performance on this particular benchmark. Future work involves evaluation of these systems on a separate balanced dataset of similar demographic attributes to PPB or making use of metrics such as balanced error to account for class imbalances in existing benchmarks.

Additionally, although Face++ appears to be the least engaged or responsive company, a limitation of the survey to English blog posts and American mainstream media quotes (Face++ 2018), definitively excludes Chinese media outlets that would reveal more about the company's response to the audit.

Conclusion

Therefore, we can see from this follow-up study that all target companies reduced classification bias in commercial APIs following the Gender Shades audit. By highlighting the issue of classification performance disparities and amplifying public awareness, the study was able to motivate companies to prioritize the issue and yield significant improvements within 7 months. When observed in the context of non-target corporation performance, however, we see that significant subgroup performance disparities persist. Nevertheless, corporations outside the scope of the study continue to speak up about the issue of classification bias (Brackeen 2018b). Even those less implicated are now facing increased scrutiny by civil groups, governments and the consumers as a result of increased public attention to the issue (Snow 2018). Future work includes the further development of audit frameworks to understand and address corporate engagement and awareness, improve the effectiveness of algorithmic audit design strategies and formalize external audit disclosure practices.

Furthermore, while algorithmic fairness may be approximated through reductions in subgroup error rates or other performance metrics, algorithmic justice necessitates a transformation in the development, deployment, oversight, and regulation of facial analysis technology. Consequently, the potential for weaponization and abuse of facial analysis technologies cannot be ignored nor the threats to privacy or breaches of civil liberties diminished even as accuracy disparities decrease. More extensive explorations of policy, corporate practice and ethical guidelines is thus needed to ensure vulnerable and marginalized populations are protected and not harmed as this technology evolves.

References

Allen D. Householder, Garret Wassermann, A. M., and King, C. 2017. The cert guide to coordinated vulnerability disclosure. Government technical report, Carnegie Mellon University.

Aniko Hannak, Claudia Wagner, D. G. A. M. M. S., and Wilson, C. 2017. Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In *2017 ACM Conference*, 1914–1933. New York, NY, USA: ACM.

Arjuna Capital, As You Sow, C. A. M. J. C. D. S. o. H. D. I. L. F. . I. S. F. H. P. C. H. I. I. M. S. M. N. C. f. R. I. S. A. M. T. S. E. G. T. S. G. o. L. W. . C. T. W. M. L. U. S. o. T. U. P. W. A. M. Z. A. M. 2018. Letter from shareholders to amazon ceo jeff bezos regarding rekognition.

Brackeen, B. 2018a. Face off: Confronting bias in face recognition ai.

Brackeen, B. 2018b. Facial recognition software is not ready for use by law enforcement.

Brendan F. Klare, Mark J. Burge, J. C. K. R. W. V. B., and Jain, A. K. 2012. Face recognition performance: Role of demographic information. In *IEEE Transactions on Information Forensics and Security*, volume 7, 1789–1801. New York, NY, USA: IEEE.

Buolamwini, J., and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of Machine Learning Research*. Conference on Fairness, Accountability, and Transparency.

Buolamwini, J. 2017. Gender shades.

Buolamwini, J. 2018. When the robot doesnt see dark skin.

Burrell, J. 2016. How the machine thinks: Understanding opacity in machine learning algorithms. *Big Data & Society*.

Cagle, M., and Ozer, N. 2018. Amazon teams up with government to deploy dangerous new facial recognition technology.

Christian Sandvig, Kevin Hamilton, K. K., and Langbort, C. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. Data and Discrimination, Converting Critical Concerns into Productive: A Preconference at the 64th Annual Meeting of the International Communication Association.

Coldewey, D. 2017. Sen. harris tells federal agencies to get serious about facial recognition risks.

Crenshaw, K. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum* 1989(8).

Diakopoulos, N. 2016. Accountability in algorithmic decision making. *Communications of the ACM* 59(2):56–62.

Edelman, B., and Luca, M. 2014. Digital discrimination: The case of airbnb.com. *SSRN Electronic Journal*.

Eslami, M.; Vaccaro, K.; Karahalios, K.; and Hamilton, K. 2017. Be careful, things can be worse than they appear: Understanding biased algorithms and users' behavior around them in rating platforms. In *ICWSM*.

Face++. 2018. Notice: newer version of face detect api.

Gary Soeller, Karrie Karahalios, C. S., and Wilson, C. 2016. Mapwatch: Detecting and monitoring international border personalization on online maps. *J. ACM*.

IBM. 2018. Ibm principles for trust and transparency.

Irene Chen, F. D. J., and Sontag, D. 2018. Why is my classifier discriminatory? In *arXiv preprint*. arXiv.

Jakub Mikians, Lszl Gyarmati, V. E. a. N. L. 2012. Detecting price and search discrimination on the internet. In *Proceedings of the 11th ACM Workshop on Hot Topics in Networks, HotNets-XI*. New York, NY, USA: ACM.

Jonathon Phillips, Fang Jiang, A. N. J. A., and OToole, A. J. 2011. An other-race effect for face recognition algorithms. In *ACM Transactions on Applied Perception (TAP)*, volume 8. ACM Press.

Juhi Kulshrestha, Motahhare Eslami, J. M. M. B. Z. S. G. K. P. G. K. K. 2017. Quantifying search bias: Investigating sources of bias for political searches in social media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 417–432. New York, NY, USA: ACM.

Julia Angwin, Jeff Larson, S. M., and Kirchner, L. 2016. Machine bias.

Kevin Hamilton, Karrie Karahalios, C. S., and Eslami, M. 2014. A path to understanding the effects of algorithm awareness. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems (CHI EA '14)*, 631–642. New York, NY, USA: ACM.

Kevin Hamilton, Motahhare Eslami, A. A. K. K., and Sandvig, C. 2015. I always assumed that i wasn't really that close to [her]: Reasoning about invisible algorithms in the news feed. In *Proceedings of 33rd Annual ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM.

Le Chen, A. M., and Wilson, C. 2015. Peeking beneath the hood of uber. In *Proceedings of 2015 ACM Conference*, 495–508. New York, NY, USA: ACM.

Mei Ngan, M. N., and Grother, P. 2015. Face recognition vendor test (frvt) performance of automated gender classification algorithms. Government technical report, US Department of Commerce, National Institute of Standards and Technology.

Philip Adler, Casey Falk, S. A. F. T. N. G. R. C. S. B. S., and Venkatasubramanian, S. 2018. Auditing black-box models for indirect influence. *Knowledge and Information Systems* 54(1).

P.J. Phillips, Hyeonjoon Moon, S. R., and Rauss, P. 2000. The feret evaluation methodology for face-recognition algorithms. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 22, 1090 – 1104. IEEE.

Puri, R. 2018. Mitigating bias in ai models.

Riccardo Guidotti, Anna Monreale, F. T. D. P., and Giannotti, F. 2018. A survey of methods for explaining black box models. *ACM Computing Surveys* 51(5).

Roach, J. 2018. Microsoft improves facial recognition technology to perform well across all skin tones, genders.

Sandra Wachter, B. M., and Russell, C. 2018. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law & Technology* 31(2).

Smith, B. 2018. Facial recognition technology: The need for public regulation and corporate responsibility.

Snow, J. 2018. Amazon's face recognition falsely matched 28 members of congress with mugshots.

Tramèr, F.; Atlidakis, V.; Geambasu, R.; Hsu, D. J.; Hubaux, J.-P.; Humbert, M.; Juels, A.; and Lin, H. 2015. Discovering unwarranted associations in data-driven applications with the fairtest testing toolkit. *CoRR* abs/1510.02377.