

MIT Open Access Articles

Actionable Auditing, Revisited

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation: Raji, Inioluwa and Buolamwini, Joy. 2022. "Actionable Auditing, Revisited."

As Published: <https://doi.org/10.1145/3571151>

Publisher: ACM|Communications of the ACM

Persistent URL: <https://hdl.handle.net/1721.1/147670>

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Terms of use: Creative Commons Attribution 4.0 International license



Actionable Auditing Revisited—

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

By Inioluwa Deborah Raji and Joy Buolamwini

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 and continue to find it difficult to translate such independent assessments into meaningful corporate accountability. 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 such as 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 such as 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 application program interface (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 5.72–8.3% 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. This is an expanded version of an earlier publication of these results, revised for a more general audience, and updated to include commentary on further developments.

1. 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.

However, researchers who engage in algorithmic audits regularly face certain difficulties. Given these risks, much algorithmic audit work has focused on goals to gauge *user* awareness of algorithmic bias⁷ or evaluate the impact of bias on *user* behavior and outcomes,^{6,15} rather than directly challenging companies to change commercial systems. Research on the real-world impact of an algorithmic audit for corporate accountability purposes is thus needed to inform strategies on how to engage corporations productively in addressing algorithmic bias and inform algorithmic auditing practices that lead to tangible accountability outcomes.

This paper maps out the real-world consequences of the implementation of an “actionable audit,” that is, an algorithmic audit designed to galvanize corporate action. In particular, the Buolamwini & Gebru Gender Shades study,³ which investigated the accuracy of commercial gender classification services, provides an apt case study. As demonstrated in a 2019 National Institute of Standards and Technology (NIST) study,⁹ the bias concerns raised by the Gender Shades study regarding gender classification also apply to facial recognition applications more broadly. Thus, the audit not only revealed the biased performance of commercial facial analysis products, but also spurred a highly visible corporate and public response regarding concerns for the functionality of various facial recognition applications, informing key product updates, interventions to corporate practice, and effective policy campaigns.

This paper is an extended and revised version of a past publication of these results.²⁰

The original version of this paper was published in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, January 27–28, 2019, Honolulu, HI, USA.

2. RELATED WORK

2.1. Corporate accountability and algorithms

Targeted public algorithmic audits provide one mechanism to incentivize corporations to address algorithmic bias in their offerings. Corporate accountability is understood as a company's perceived obligation to be accountable to all stakeholders affected by its activities and outputs, including but not restricted to its shareholders or users. However, outside of the capitalist motivations of economic benefit, employee satisfaction, competitive advantage, public reputation, and recent legal developments such as the EU General Data Protection Regulation, corporations still have little incentive to disclose additional details about their systems in order to be properly held to account. For instance, while corporations have a business interest in keeping their models and data private, civil society has a public interest in assuring equitable treatment of all members of society who interact with any seriously consequential product. Thus, external pressure remains a necessary approach to increase transparency and address harmful model bias. Unlike internal audit practice, executed by employees within companies as a compliance or quality control measure,²³ *external algorithmic audits* are necessary to allow for independent scrutiny of corporate products in the context of the needs of stakeholders that may seem removed from immediate corporate priorities or perspectives.

2.2. Black box algorithmic audits

The algorithmic audit has become increasingly formalized as a strategy to expose systematic biases embedded in software platforms. A dominant understanding involves the close analogy to the social scientific audit, where the mechanism of the algorithm or consent of the audit target is not required to evaluate the algorithm's behavior, as outcomes are collected and analyzed through the stimulation of a mock user population.⁴ Commercial application program interfaces (APIs), the focus of this study, differ from these formulations as they are platforms that are selected and used by intermediary stakeholders (i.e., developers) separated from the users that may be subject to the biased outcomes in a final product. Thus, the simulation of a developer rather than user population is required and provides an opportunity to reconsider the formulation of end user platform 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 test sets such as FERET and the Facial Recognition Vendor Test (FRVT) from the National Institute of Standards and Technology (NIST) is of particular interest, as examples specific to establishing policy and legal restrictions around mitigating bias in facial recognition technologies. Such vendor tests mainly involve an analysis of a model's performance on a given benchmark dataset, without requiring anything beyond user-level access to the model itself to determine biased performance outcomes.

3. GENDER SHADES

The Gender Shades study differs from these previous cases as an external and multi-target black box audit of commercial machine learning APIs, scoped to evaluating the facial analysis task of binary gender classification.³

The contribution of the work is twofold, 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 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 such as darker female, darker male, lighter female, and lighter male.

3.1. Gender Shades audit design

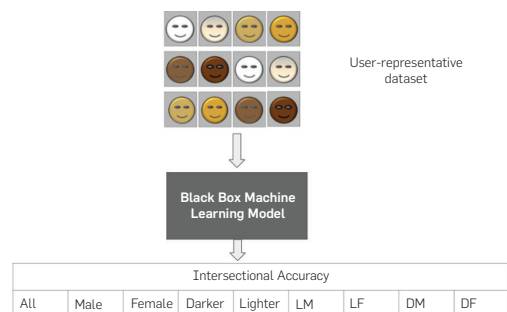
In this section, we articulate the identifiable structure and procedure design of the Gender Shades study, to identify the elements employed that led to more effective corporate responses.

3.2. Pilot Parliaments Benchmark

A benchmark is the dataset used in the evaluation of a model. There has been a long-standing issue of *benchmark bias* in algorithmic development—where failures on specific subpopulations that may be impacted by the model's deployments are not part of these test datasets. As these subjects do not have their data considered in evaluations, there is little to no understanding of model performance on the population that is underrepresented or missing in the benchmark data.

The Gender Shades paper contributed the Pilot Parliaments Benchmark (PPB) to the computer vision community. 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

Figure 1. Gender Shades audit process overview.³



the diversity of that group. With more 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. Algorithmic unfairness is evaluated by comparing classification accuracy across identified subgroups in the user-representative test set (see Figure 1).

Although prior studies have looked at gender and racial discrimination in facial recognition separately,^{13,14} *intersectional* considerations remain lacking—particularly in the scenario similar to Gender Shades where gender, a protected attribute, is also the predicted class and cannot be obscured or minimized without influencing model performance. Compared with other mainstream face datasets at the time, the PPB was more balanced with respect to both representation for skin type and gender—the benchmark is 55.4% male and 44.6% female faces as well as containing 46.4% darker- and 53.6% lighter-skinned subjects.³ Comparatively, other mainstream face datasets were as little as 22.5% female and 77.4% male, with as few as 5.4% darker-skinned subjects and up to 94.6% lighter-skinned subjects.¹⁷

3.3. Named multitarget audit

It can be difficult to get the public or other external stakeholders to pay attention to audit results and become aware of the findings. However, engaging public participation and *enciting public pressure* are necessary for external accountability processes.

In several implemented audit studies, vendor names are kept anonymous¹⁴ or the scope is scaled down to a single named target.¹⁵ The former fails to increase public awareness of a particular target, and the latter fails to capture the competitive dynamics of a named multi-target audit—thus reducing the impetus for corporate reactions to those studies.

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.¹¹ In a similar style to such audits, the audited corporations are thus *named* explicitly and the products of *multiple* companies in the industry are targeted. The result of this is a competitive dynamic due to mutual awareness of audit results—if one corporation reacts, the others that are named will feel pressured to do so as well, to remain competitive in the market shared with other targets and avoid further scrutiny. The naming of multiple audit targets also increases the likelihood of interest in audit results from the public and the press, influencing the desired public pressure required to make audit results more influential.

3.4. Carrier puppet audit design

Another challenge to audit work is *access*. It can be difficult to determine when biased outcomes arise as the direct result of a model (rather than some other factor of an application’s design). Furthermore, auditors run the risk of breaching company Terms of Service, the Computer Fraud and Abuse

Act (CFAA), or ACM ethical practices when they interact with products beyond the advertised use case.

Thus, a key element of Gender Shade’s audit design is that the audit targets are commercial machine learning APIs. The auditors thus mirror the behavior 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 model⁴ 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 toward end users (see Figure 2).

By puppeting the developer user, auditors target the API directly for the audit, rather than the end application—this design decision lets auditors access the model directly while still operating at a consumer level of access.

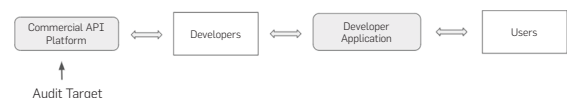
3.5. Gender Shades-coordinated bias disclosure

Communication with corporations around potential system limitations, vulnerabilities, or failures has historically been a delicate issue. Those conducting audits face much uncertainty around *hostile corporate reactions*.

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, 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, with respect to general public awareness and careful to guard against corporate inaction.¹¹ 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.

Gender Shades auditors approached audited corporations systematically, following a procedure sequentially outlined below that closely mirrors key recommendations for coordinated vulnerability disclosures (CVDs) in information security.¹¹

Figure 2. “Carrier puppet” audit framework overview.



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.³
2. **Defined corporate response period with limited anonymized release to audit targets**—The Gender Shades paper (without explicit company references) was sent to Microsoft, IBM, and Face++ on December 19, 2017,² 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 9, 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 study² 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 prompt 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 24, 2018, with explicit company references at the FAT* conference to an audience of academics, industry stakeholders, and policymakers.²
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 of Gender Shades, the lead author presented and linked to IBM’s updated API results at the time of the public release of the initial study.²

4. METHODOLOGY

Now that we have done an overview of the key design considerations that informed Gender Shades, we assess the study’s effectiveness in leading to tangible updates of corporate performance on the audit benchmark. To do so, we first re-conduct an audit similar to that initially implemented in the Gender Shades study a year prior. We reimplement this audit for the same products evaluated in the initial study (i.e., target corporations) and also a set of products from companies not included in the initial study (i.e., non-target corporations). We then discuss and contextualize any performance changes with additional information about corporate communications and policy developments over the intervening period between this audit and the initial Gender Shades study. In this period, the PPB benchmark had not been shared publicly or with any corporate entity.

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 market share, the availability of desired API functions, and overall market influence as driving factors in the decision to select Microsoft, IBM, and Face++.³ 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.¹ Non-target corporation Amazon was selected following the revelation of the active use and promotion of its facial recognition technology in law enforcement.²⁵

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 dataset $D = (X, Y, C)$, 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:

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

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.¹ 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.

5. PERFORMANCE RESULTS

With the results of the follow-up audit and original Gender Shades outcomes, we first analyze the differences between the performances 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.

The reported follow-up audit was done on August 21, 2018, for all corporations in both cases. Summary of Tables 1 and 2 shows 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.

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

5.1. Key findings of target corporations

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,⁸ while IBM was the first to release a new API version in 66 days,¹⁹ with Microsoft updating their product the day before Face++, in 189 days.²⁴ All targeted classifiers in post-audit releases have their largest error rate for the darker female subgroup and the lowest error rate for the lighter male 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.

5.2. Key findings of non-target corporations

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 Microsoft’s 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 Gebru³ and Mei Ngan and Grother.¹⁶
- Kairos and Amazon perform better on lighter faces than darker faces, a trend also observed in Buolamwini and Gebru³ and Jonathon Phillips et al.¹³
- 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 the highest and lowest accuracy intersectional subgroups, while Amazon has an error gap of 31.37%.

6. 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.

6.1. 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,⁵ 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 refer Gender

Shades in product update releases, claim intersectional model improvements on their gender classifier.^{19, 24} 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 5.72–8.3% 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.

However, we must be careful not to assume that reductions in subgroup accuracy disparities for gender classification indicate performance improvements across all facial recognition tasks. Following the publication of these findings, a National Institute of Standards and Technology (NIST) study evaluated 189 software algorithms—a majority of the industry.⁹ The study found that for the facial recognition task of verification (i.e., one-to-one matching), the false-positive rates for Asian and African-American faces were 10–100 times higher than this error rate for Caucasian faces, depending on the individual algorithm. In particular, the NIST team regularly observed the highest rates of false positives for African-American females, indicating a greater risk of misidentification for this group. This indicates that these functionality issues are relevant across a range of facial recognition tasks and that technical updates on one task do not necessarily translate to updates on addressing disparate performance on another task.

6.2. Corporate prioritization

Although the original study³ 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-aware corporations can also drastically alter the processes needed to reduce bias in algorithmic systems.

6.3. 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,^{1, 19, 24} 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 research¹² 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.

As discussed in follow-up work,²² the act of diversifying datasets can come with serious trade-offs with respect to privacy risks, especially if the data involved are an identifiable biometric, as is the case with facial recognition. Consent violations during the data collection process and inadequate data storage or dissemination practices can lead to a set of additional issues. For instance, IBM built the Diversity in Faces dataset to address the functional limitations addressed in the Gender Shades audit. This dataset was sourced from Creative Commons licensed images uploaded to Flickr by users who could not possibly consent to being included in a facial recognition dataset.¹⁷ In fact, the history of facial recognition reveals consistent issues with proper data management,²¹ reflecting a broader negligence of conscientious data practices in the machine learning field.¹⁸

6.4. 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 30, 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 29, 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.

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.^{19,24}

Only Microsoft strongly mentions the solution of Diversity & Inclusion considerations in hiring as an avenue to address issues.²⁴ 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”.¹ A cultural shift in the facial recognition industry could thus attract and retain those paying increased attention to the issue due to personal resonance.

6.5. 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 non-target product improvements, Kairos and Amazon both perform more closely to the target corporations’ performance in the Gender Shades study than their performance in this follow-up audit.

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,²⁵ 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.³

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.

6.6. 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 referenced 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. Similar calls for action to Axon AI by several civil rights groups, as well as letters from then Senator, now Vice President Kamala D. Harris to the EEOC, FBI, and FTC, regarding the use of facial recognition in law enforcement also directly reference the work.

Several states in the United States—California, Washington, Idaho, Texas, and Illinois—in addition to some cities—San Francisco, Oakland, and Somerville—are already taking the lead in regulating or outright banning the use of these technologies through coordinated campaigns such as the ACLU’s Community Control Over Police Surveillance (CCOPS) initiative. As of the writing of this paper, federal

bill proposals such as the Algorithmic Accountability Act, Commercial Facial Recognition Privacy Act of 2019, No Biometric Barriers Act, and the Facial Recognition and Biometric Technology Moratorium Act of 2020 have also been proposed in the U.S.

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. In fact, Microsoft goes so far as to explicitly support public regulation. Following the widespread protests for racial equity in the United States in June 2020, IBM, Microsoft, and Amazon all made varying degrees of commitments to a voluntary recall of their facial recognition products, all announcing an intent to stop the sale of their products to police clients. IBM pulled out of the market completely, terminating their API endpoint and declaring an exit from the facial recognition market. Amazon made a brief statement of pausing police use of their facial recognition products for at least a year, and Microsoft reiterated commitments to not allow for the use of their facial recognition products by police until “the government passes federal legislation regulating the technology.”¹⁰ This indicates an increasing industry-wide acknowledgment of the limitations of these products and the practical risks involved in deploying a product with such consequential flaws and biases.

7. 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.³ 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 females¹⁹—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.³ 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,⁸ it definitively excludes Chinese media outlets that would reveal more about the company’s response to the audit.

Since the initial publication of this study,²⁰ the limitations

of narrow scope algorithmic audits like Gender Shades have been further scrutinized.²² This narrow scope can facilitate greater impact, focusing on efforts of improvement on addressing the highest risk threats. However, doing so also significantly limits the scope of the audit's impact and can lead to an over-emphasis on technical improvements that overfit to the specified demographic groups, prediction tasks, or companies included in a particular audit.

8. CONCLUSION

We can see from this follow-up study that the Gender Shades audit was able to motivate target companies to prioritize addressing classification bias in their systems 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.¹ 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.²⁵ 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 are thus needed to ensure that vulnerable and marginalized populations are protected and not harmed as this technology evolves. 

References

- Brackeen, B. Facial recognition software is not ready for use by law enforcement, 2018.
- Buolamwini, J. Gender shades, 2017.
- Buolamwini, J., Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of Machine Learning Research*. Conference on Fairness, Accountability, and Transparency, February 2018.
- Christian Sandvig, K.K., Hamilton, K., Langbort, C. Auditing algorithms: Research methods for detecting discrimination on internet platforms. Data and Discrimination, Converting Critical Concerns into Productive: A Pre-conference at the 64th Annual Meeting of the International Communication Association, 2014.
- Crenshaw, K. 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), 1989.
- Edelman, B., Luca, M. Digital discrimination: The case of airbnb.com. *SSRN Electron. J* 2014.
- Eslami, M., Vaccaro, K., Karahalios, K., Hamilton, K. Be careful, things can be worse than they appear: Understanding biased algorithms and users' behavior around them in rating platforms. In *ICWSM*, 2017.
- Face++. Notice: newer version of face detect api, 2018.
- Grother, P., Ngan, M., Hanaoka, K. *Face Recognition Vendor Test (FVRT): Part 3, Demographic Effects*. National Institute of Standards and Technology, 2019.
- Heilweil, R. Big tech companies back away from selling facial recognition to police that's progress. *Recode*, 2020. <https://www.vox.com/recode/2020/6/10/21287194/amazon-microsoft-ibm-facial-recognition-moratorium-police>.
- Householder, A.D., Wassermann, G., King, C. The cert guide to coordinated vulnerability disclosure. Government Technical Report, Carnegie Mellon University, 2017.
- Irene Chen and D. Sontag. Why is my classifier discriminatory? In *arXiv preprint*. arXiv, 2018.
- Jonathon Phillips, A.N.J.A., Jiang, F.,

- O'Toole, A.J. An other-race effect for face recognition algorithms. In *ACM Transactions on Applied Perception (TAP)*. Volume 8. ACM Press, 2011.
- Klare, B.F., Burge, M.J., Jain, A.K. Face recognition performance: Role of demographic information. In *IEEE Transactions on Information Forensics and Security*. Volume 7. IEEE, New York, NY, USA, 2012, 1789–1801.
- Kulshrestha, J., Eslami, M. 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*, ACM, New York, NY, USA, 2017, 417–432.
- Mei Ngan, M.N., Mei NganGrother, P. Face recognition vendor test (frvt) performance of automated gender classification algorithms. Government Technical Report, US Department of Commerce, National Institute of Standards and Technology, 2015.
- Merler, M., Ratha, N., Feris, R.S., Smith, J.R. Diversity in faces. *arXiv preprint arXiv:1901.10436*, 2019.
- Paullada, A., Raji, I.D., Bender, E.M., Denton, E., Hanna, A. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *arXiv preprint arXiv:2012.05345*, 2020.
- Puri, R. Mitigating bias in ai models, 2018.
- Raji, I.D., Buolamwini, J. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, 429–435.
- Raji, I.D., Fried, G. About face: A survey of facial recognition evaluation. *arXiv preprint arXiv:2102.00813*, 2021.
- Raji, I.D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., Denton, E. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, 145–151.
- Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., mith-Loud, J., Theron, D., Barnes, P. Closing the ai accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, 33–44.
- Roach, J. Microsoft improves facial recognition technology to perform well across all skin tones, genders, 2018.
- Snow, J. Amazon's face recognition falsely matched 28 members of congress with mugshots, 2018.

Inioluwa Deborah Raji ([deborah.raji]@mail.utoronto.ca), University of Toronto, Canada.

Joy Buolamwini ([joyab]@mit.edu), Massachusetts Institute of Technology, Cambridge, MA, USA.



This work is licensed under a Creative Commons Attribution International 4.0 License. <https://creativecommons.org/licenses/by/4.0/>