Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing

Inioluwa Deborah Raji
University of Toronto
Canada
deborah.raji@mail.utoronto.ca

Timnit Gebru
Google Research
United States
tgebru@google.com

Margaret Mitchell
Google Research
United States
mmitchellai@google.com

Joy Buolamwini
MIT Media Lab.
United States
joy.buolamwini@gmail.com

Joonseok Lee
Google Research
United States
joonseok@google.com

Emily Denton
Google Research
United States
dentone@google.com

ABSTRACT

Although essential to revealing biased performance, well intentioned attempts at algorithmic auditing can have effects that may harm the very populations these measures are meant to protect. This concern is even more salient while auditing biometric systems such as facial recognition, where the data is sensitive and the technology is often used in ethically questionable manners. We demonstrate a set of five *ethical concerns* in the particular case of auditing commercial facial processing technology, highlighting additional design considerations and ethical tensions the auditor needs to be aware of so as not exacerbate or complement the harms propagated by the audited system. We go further to provide tangible illustrations of these concerns, and conclude by reflecting on what these concerns mean for the role of the algorithmic audit and the fundamental product limitations they reveal.

ACM Reference Format:
Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20), February 7–8, 2020, New York, NY, USA*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3375627.3375820

1 INTRODUCTION

Facial processing technology (FPT) is a broad term that encompasses a variety of tasks ranging from face detection, which involves locating a face within a bounding box in an image; facial analysis, which determines an individual's facial characteristics including physical or demographic traits; and face verification or identification, which is the task of differentiating a single face from others.

FPT can be deployed for a wide range of uses ranging from smiling detection to gauge customer satisfaction, to estimating the demographic characteristics of a subject population, to tracking individuals using face identification tools [37]. Corporate rhetoric on the positive uses of this technology includes claims of "understanding users", "monitoring or detecting human activity", "indexing and searching digital image libraries", and verifying and identifying subjects "in security scenarios" [1, 8, 16, 33].

The reality of FPT deployments, however, reveals that many of these uses are quite vulnerable to abuse, especially when used for surveillance and coupled with predatory data collection practices that, intentionally or unintentionally, discriminate against marginalized groups [2]. The stakes are particularly high when we consider companies like Amazon and HireVue, who are selling their services to police departments or using the technology to help inform hiring decisions respectively [5, 43].

Civil rights organizations have already sounded the alarm against facial recognition technology in particular and the need for urgent policy and regulatory action to restrict its use. 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 [9], Commercial Facial Recognition Privacy Act of 2019 [4] and No Biometric Barriers Act [10] have also been proposed in the U.S., as well as a bill proposing a moratorium in the U.K. [11].

Several of these proposals and their corresponding memos explicitly recommend that the results of FPT audits such as Gender Shades [6] and benchmarks developed through the National Institute of Standards and Technology [38] serve as conditions for FPT accreditation or moratorium. The language used in these proposals frames such audits as trusted mechanisms to certify the technology as safe and reliable for deployment.

In this paper, we caution against this stance, outlining ethical concerns we have identified in the development and use of these algorithmic audits. We believe these concerns to be inherent restrictions to the utility of these audits within the broader evaluation of these systems, and propose to explicitly acknowledge these limitations as we make use of the audits in practice and in policy.

Our primary contributions are as follows. We first develop CelebSET, a new intersectional FPT benchmark dataset consisting of celebrity images, and evaluate a suite of commercially available FPT APIs using this benchmark. We then use our benchmark and

AIES '20, February 7–8, 2020, New York, NY, USA

ACM ISBN 978-1-4503-7110-0/20/02.
https://doi.org/10.1145/3375627.3375820

audit development process as a case study, and outline a set of ethical considerations and ethical tensions relevant to algorithmic auditing practices.

2 CELEBSET: A TYPICAL FPT BENCHMARK

Datasets such as Face Recognition Vendor Tests (FRVT), IARPA Janus Benchmarks (IJB) [38], and the Pilot Parliaments Benchmark (PPB) [6] have been key to identifying classification bias in the specific tasks of face detection, verification, identification, identity clustering and gender recognition. Audits conducted using these datasets have heavily informed many of the recent attempts to address the dearth of diversity in FPT benchmarks and proposed frameworks for evaluating classification bias in deployed systems [32, 43, 45]. Although these prior studies evaluate commercially deployed models for gender and racial discrimination in FPT, there are few benchmarks that enable the evaluation of the full range of classification tasks commercially available in facial processing technology today, and none that enables an intersectional black box audit evaluation for such a wide range of tasks.

To address this gap, we develop a benchmark called CelebSET, a subset of the IMDB-WIKI dataset [40] that includes 80 celebrity identities–20 of the most photographed celebrities from each of the subgroups 'darker male'(DM), 'darker female'(DF), 'lighter male'(LM) and 'lighter female'(LF). Metadata pertaining to celebrity ethnicity were crawled from the celebrity fan websites FamousFix.com and ethiccelebs.com, and indexed to celebrity identities from the IMDB-WIKI dataset.'Darker'(D) is approximated by taking a subset of the images of celebrities tagged with the crawled 'Black' label from our crawled dataset, and selecting subjects within the 3 to 6 range of the Fitzpatrick skin type scale [41] using the reported Fitzpatrick labels of celebrity references for guidance. The 'Lighter'(L) individuals are a subset of subjects with the crawled 'White' label, also verified through visual inspection to approximate the skin type of celebrity references for a Fitzpatrick scale of 1 to 3. Gender - defined as 'Male'(M) and 'Female'(F) - as well as age metadata is taken from the IMDB-WIKI dataset. We estimate the age of the celebrity depicted in each photo by subtracting the person's known date of birth from the timestamp of the uploaded photo. We manually identify 10 smiling and 10 non-smiling example images for each celebrity.

We use the original uncropped images with bounding box labels provided from the IMDB-WIKI dataset [40] to perform our audit on the detection task and use cropped face images from the original dataset to audit facial analysis tasks. The full dataset with metadata, cropped and uncropped images is available as Supplementary Materials.

2.1 API Evaluation on CelebSET

We evaluate the APIs of Microsoft, Amazon, and Clarifai, which offer the widest scope of facial analysis tasks. Microsoft is notable as a target corporation in the initial Gender Shades study [6], and Amazon in the follow up study, Actionable Auditing [39]. Clarifai is notable as an API that has not been previously included in any prior audit studies. For tasks such as automatic gender recognition, smile detection, and name identification, we evaluate the accuracy

Table 1: Overall accuracy on designated facial analysis prediction tasks.

	Gender	Age	Name	Smile	Detection
Microsoft	99.94%	74.09%	98.69%	79.94%	93.56%
Amazon	99.75%	58.40%	87.25%	94.16%	99.25%
Clarifai	85.97%	55.24%	95.00%	56.19%	99.31%

Table 2: Difference in accuracy between the lighter (L) subgroup and darker (D) subgroup for each prediction task.

Task	Gender	Age	Name	Smile	Detection
Microsoft	0.13%	18.35%	1.41%	-0.48%	3.38%
Amazon	0.25%	16.83%	1.03%	-0.75%	0.25%
Clarifai	11.69%	1.00%	7.50%	0.12%	0.42%

Table 3: Difference in accuracy between the Male (M) subgroup and female (F) subgroup for each prediction task.

Task	Gender	Age	Name	Smile	Detection
Microsoft	0.13%	9.90%	1.23%	-4.45%	0.62%
Amazon	0.00%	12.28%	4.75%	-9.00%	0.50%
Clarifai	7.58%	10.26%	-1.01%	1.25%	-1.63%

of the predicted value as compared to ground truth. For age prediction, we allow for an 8-year acceptance margin to accommodate the age range results of the Amazon API. We evaluate detection performance using the AP_{50}, the average precision at a threshold of 0.50 intersection over union (IoU). All evaluation results are as of October 2019. Calculation details, code and complete results of the audit are included in Supplementary Materials.

2.1.1 Overall Performance. As shown in Table 1, all APIs perform best on the task of gender classification–with the exception of Clarifai which performs best on face detection. It is noteworthy that two of the target APIs, Amazon and Microsoft, have been publicly audited for gender classification in previous studies, and have since released new API versions in response to the audits citing improvements [6, 39]. All APIs perform worst on age classification, with Amazon and Clarifai performing slightly better than chance on this task.

2.1.2 Performance on Unitary Subgroups. Tables 2–3 show that with few exceptions, all APIs perform worst on the darker subgroup and female subgroup across tasks, a finding supported by previous work [6]. Clarifai, the only commercial API which was not previously publicly audited for the task of gender classification, demonstrated notably higher disparities across unitary groups for that task compared to the competing APIs.

2.1.3 Performance on Intersectional Subgroups. Table 4 lists intersectional subgroup performance and shows patterns found in previous work: that the most common "least accurate subgroup" is the darker female subgroup and the most common "most accurate subgroup" is the lighter male subgroup, though there are varied configurations with exceptions to this trend.

Table 4: Difference in accuracy between the best and worst performing intersectional subgroups by prediction task. The subgroups are darker females (DF), darker males (DM), lighter females (LF) and lighter males (LM). Values in bold denote equal performance. For instance, 0.25% (DM/LM/LF - DF) signifies that the difference in accuracy between DM and DF, i.e. DM-DF, LM-DF and LF-DF are all 0.25%.

	Gender	Age	Name	Smile	Detection(AP_{50})
Microsoft	0.25% **(DM/LM/LF - DF)**	29.47% (LF-DF)	3.90% (LF-DF)	8.02% (LF-LM)	4.25% (LM-DM)
Amazon	0.50% (LF-DF)	29.10%(LM-DF)	6.71% (DM-DF)	9.75% (DF-LM)	0.75% **(LM-DF/LF)**
Clarifai	19.10% (LM-DF)	11.21% (LM-DF)	10.50% (LM-DF)	3.00% (LF-LM)	0.50% **(LM/LF-DF)**

3 ACKNOWLEDGING AND WORKING THROUGH AUDIT ETHICAL CONCERNS

Using our audit conducted with CelebSET as an example, we walk through ethical concerns in current algorithmic auditing practices. We separate these concerns into design considerations and tensions. While ethical *design considerations* outline additional points to be noted during the audit design in order for the audit to truthfully represent the performance of the system, ethical *tensions*, represent situations where different ethical ideals come into conflict and hard decisions need to be made regarding an appropriate path forward.

3.1 Design Considerations

3.1.1 Consideration 1: Selecting Scope of Impact. Algorithmic audits can target a specific demographic group, prediction task, or company. This narrow scope of targets can facilitate greater impact, focusing efforts of improvement on addressing the highest risk threats. However, doing so also significantly limits the scope of the audit's impact, and allow institutions to overfit improvements to the specified tasks.

The practical reliability of the results of a benchmark also depends on the contextual and temporal relevance of the data used in evaluations to the audit use case. If it is not communicated when it is appropriate to use a benchmark, then there is no indication of when it becomes an obsolete measure of performance. This also applies for aligning the context of use of the audited system and the audit - if one demographic is under-represented in a benchmark, then it should not be used to evaluate a model's performance on a population within that demographic. Even with intersectional considerations, there is a limit to the scope of which categories are included.

Illustration The audits conducted through CelebSET reveal that these types of external audits can only be used as an accountability mechanism within a narrow scope of influence. For instance, Clarifai has a 19.10% discrepancy between its best performing subgroup (lighter male) and worst performing subgroup (darker female) for the gender classification task, mirroring results from the original audits of the Gender Shades study [6] and demonstrating a much greater disparity compared to the difference in error rates for Microsoft and Amazon (0.25% and 0.50% respectively). In fact, both Amazon and Microsoft have their lowest intersectional discrepancies in gender classification, a task they have been both publicly audited for, and for which both companies released updated APIs after the disclosure of audit results. This replicates findings from the Actionable Auditing study–those that have been previously audited have smaller disparities on CelebSET, compared to those that have not been previously auditied, and thus classification bias continues to be a persistent challenge within the industry [39].

We found that this result holds not only for the audited company but also the task itself. In our CelebSET audit, we observed the largest difference in accuracy for Microsoft and Amazon is for the age classification task (a 29.47% and 29.10% discrepancy respectively between the error rates for the best performing (lighter female, and lighter male, respectively) and worst performing (darker female) subgroups). Although these companies have smaller disparities in error rates for the task of binary gender classification in response to being audited, large performance disparities are identified for other tasks. This may imply that external algorithmic audits only incentivize companies to address performance disparities on the tasks they are publicly audited for.

Also, institutions strive to reduce performance disparities across subgroups that have been the focus of prior public audits (e.g. binary gender and skin type). Since many of the audited APIs are currently proposed for use by U.S. law enforcement [7], immigration [18], and military services [30], the focus on performance across skin types may make sense in order to assess the risk to people of color who are over-policed and subject to additional profiling in these scenarios. However, there are other marginalized groups or cases to consider who may be being ignored. For instance, dysfunction in facial analysis systems locked out transgender Uber drivers from being able to access their accounts to begin work [31]. These and other issues have sparked recent work to start addressing performance disparities for this specific group [42].

While it is important to strive for equal performance across subgroups in some tasks, audits have to be deliberate so as not to normalize tasks that are inherently harmful to certain communities. The gender classification task on which previously audited corporations minimized classification bias, for example, has harmful effects in both incorrect and correct classification. For example, it can promote gender stereotypes [19], is exclusionary of transgender, non-binary, and gender non-conforming individuals, and threatens further harm against already marginalized individuals [22]. Thus, minimizing performance disparities and investing in the improvement of that task specifically may not be the most ethical focus of impact.

3.1.2 Consideration 2: Auditing for Procedural Fairness. Auditing outcomes is not enough. To adequately evaluate FPTs as systems embedded in their deployed environments, we need to consider the audit as more than the final system's performance on a single benchmark. It has been well established in tax compliance that taking a procedural fairness approach to organizational audits leads

to more effective evaluations. For example, studies in Australia and Malaysia sought to dissuade companies from looking for loopholes or averting tax payment–a corporate malpractice that costs states billions of dollars each year. Instead of simply looking at whatever financial documents were submitted, auditors instead evaluated the company's adherence to a tax compliance process [17, 36], by auditing the companies' internal practices and documentation development processes. The result was that institutions audited in this manner subsequently became more compliant (i.e., paid more of their taxes), and felt less intimidated by auditors. Similar findings have been discovered in employment peer review [14] and computer information security [15], revealing that inspecting adherence to a fixed and defined process for compliance standards is just as important as the result of the compliance audit itself.

Similarly, performance disparities surfaced by FPT audits do not necessarily capture the dynamics or integrity of the engineering design processes that led to these results. In some cases, "procedural fairness" for machine learning (ML) systems involves interpretability methods that attempt to understand how a prediction is made. In the case of automated facial analysis tasks, an example is identifying image features that are most likely to influence the output, and ensuring that these features do not encode protected attributes such as race. However, such a perspective constitutes a fairly constrained view as an FPT's effect on people is not limited to its prediction. The manner in which the technology is developed (e.g. were there predatory data collection practices?), the types of tests that are performed, the documentation made available, and the guardrails that are put in place are all important considerations. FPT evaluations such as the Face Recognition Vendor Test (FRVT) from the National Institute of Standards and Technology include some version of these qualitative considerations, such as a holistic product usability test [38]. But there remains a need for a comprehensive auditing framework which takes into account the end-to-end product development and deployment process.

Illustration To demonstrate the biases baked into the model development and design process not captured by the CelebSET audit, we examine the diversity of the selected celebrity identities included in the APIs' model design for this task. Each of the audited APIs has a model that takes as its input an image, and outputs the name of the celebrity contained in the image. We can analyze the demographic distribution of the celebrities included in each API, in order to understand who product developers consider to be a celebrity, and how representative this selection is.

We estimate the full list of celebrity names used in the Microsoft classifier through a publicly released dataset from the company which includes 100,000 logged celebrity identities [21]. Clarifai gives users access to the full list of 10,000 celebrities through its API, and Amazon does not make the list of included celebrity identities available in any form. We obtain each celebrity's race by matching their identity to ethnicity labels on FamousFix.com and ethniccelebs.com. Table 5 shows the breakdown of ethnicities that are represented in Microsoft's and Clarifai's databases. While Clarifai includes many more Caucasian celebrity identities (74% of celebrity labels) than any other group, Microsoft, with 37% Caucasian, 19% Asian and 21% Black celebrity names included appears to have a more inclusive design.

Had we focused solely on the performance of these APIs on CelebSET, we would have missed this label selection bias and remained with an incomplete understanding of the design flaws that influence the APIs' performance. While comprehensive auditing frameworks examining model development and deployment processes are yet to be developed, documentation proposals such as datasheets [20], model cards [34] and factsheets [24] can encourage designers to carefully think about these processes if they are required elements of such an audit.

3.2 Ethical Tensions

3.2.1 Tension 1: Privacy and Representation. While audit benchmark datasets should reflect the populations who will be impacted by the audited technology, collecting a sufficiently large and diverse dataset can present privacy risks for the individuals represented in the dataset. Depending on data storage and dissemination policies, sensitive and biometric information may be made accessible beyond the intended auditing purpose. These risks can be further compounded by potential consent violations during the data collection process. For example, IBM's Diversity in Faces dataset was sourced from Creative Common licensed images uploaded to Flickr [32]. While these images are open for public internet use, the Flickr users who uploaded the photos, and the individuals in the photos, did not consent to being included in a facial recognition dataset [44].

Privacy and consent violations in the dataset curation process often disproportionately affect members of marginalized communities. Benchmark dataset curation frequently involves supplementing or highlighting data from a specific population that is underrepresented in previous datasets. Efforts to increase representation of this group can lead to tokenism and exploitation, compromise privacy, and perpetuate marginalization through population monitoring and targeted violence [22, 25, 35]. And the method through which companies pursue better representation can be ethically questionable. For instance, a startup signed a deal with the Zimbabwe government to harvest the faces of millions of citizens through unprecedented access to their CCTV cameras, smart financial systems, airport, railway, and bus station security, and a national facial database [23]. Without seeking the active consent of impacted individuals and working towards mutual benefit, such an act can be exploitative and tokenizing of the humans contributing to the improvement of the system for which their data is used.

Illustration CelebSET was sourced from IMDB-WIKI [40], a dataset with significant demographic bias that can be seen in the distribution of meta-data labels in Table 6 and Table 7. Certain groups are not only highly underrepresented, but there are fewer images per person from these groups. This skew is a result of media and social biases that make certain subgroups less photographed than others, and thus less likely to exist in CelebSET's source data. This lack of representation becomes even more stark when considering intersectional identities such as Black women.

Consequently, when aiming to design a "demographically balanced" benchmark, i.e., one with equal representation from the designated demographic subgroups, it is more challenging to sufficiently represent certain groups relative to others. Thus, by actively seeking to include members of underrepresented groups, the

Table 5: Breakdown of celebrity identities in commercial APIs by ethnicity.

	Asian	White	Hispanic	Black	Middle Eastern	Indian	Other/Mixed	Total
Microsoft API	7,838	15,536	10	8,816	995	1,316	7,167	41,678
	18.8%	37.3%	0.02%	21.1%	2.4%	3.2%	17.2%	100.0%
Clarifai API	172	4,861	0	534	31	125	800	6,523
	2.6%	74.5%	0.0%	8.2%	0.5%	1.9%	12.3%	100.0%

Table 6: Breakdown of IMDB-WIKI data examples by ethnicity.

	Asian	White	Hispanic	Black	Middle Eastern	Indian	Other/Mixed	Total
IMDB-WIKI-Eth	7,557	338,896	351	29,613	1,160	3,299	33,468	414,344
	1.8%	81.8%	0.1%	7.2%	0.3%	0.8%	8.1%	100.0%

Table 7: Breakdown of IMDB-WIKI data examples by gender.

	Male	Female	Unknown	Total
IMDB-WIKI	230,912	179,900	3,532	414,344
	55.7%	43.4%	0.85%	100%

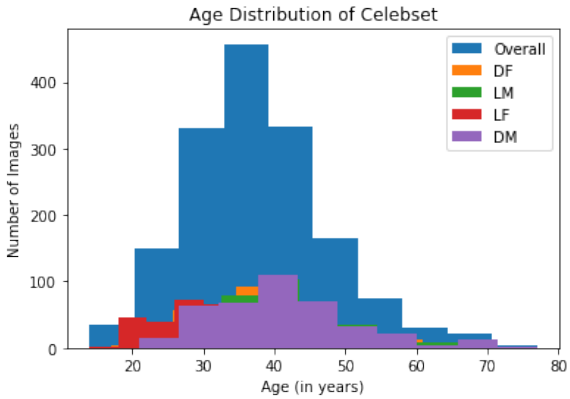

Figure 1: Histogram of the age distribution in CelebSET. The median age is 37, the mode is 36 and the mean is 37.56. The youngest subject is 14 and the oldest is 77. Blue signifies overall age distribution. Purple, orange, green and red show the age distribution for darker males, lighter males, darker females and lighter females respectively.

privacy risk is disproportionately increased for that group. For instance, in this case, there is twice the likelihood that an image from the "Black" subgroup of the reference dataset will be included in CelebSET than an image from the "White" subgroup.

In addition to the ethical challenges emerging from the skewed distribution of the data source, we also encounter challenges pertaining to obtaining consent. The CelebSET dataset is sourced from a database of public figures. While these individuals have, from a legal perspective, opted in to having their likeness used freely in the public domain, we acknowledge that they have not consented to inclusion in an FPT benchmarking dataset specifically. While an opt-in informed consent process would be ideal for subjects included in the benchmark, the individuals in CelebSET are effectively unreachable and thus cannot be contacted to give informed opt-in consent. Even the less ideal opt-out model is challenging to implement as many subjects may never become aware that their face is in the dataset, thus rendering the option to opt-out meaningless. The consistency of the benchmark can also be compromised if the dataset changes over time through the removal of individuals.

3.2.2 Tension 2: Intersectionality and Group-Based Fairness. The concept of intersectionality, coined by legal scholar Crenshaw [12], is a framework for understanding how interlocking systems of power and oppression give rise to qualitatively different experiences for individuals holding multiply marginalized identities [13].

Crenshaw writes of moving beyond stereotypical assignments and recognizing decision outcomes on a more individualized basis – observing an individual to be at the intersection of numerous unique combinations of identities, possessing several dimensions of privilege and oppression. However, in order for group fairness to be evaluated, an individual's experience must be reduced to a categorical assignment, even while performing disaggregated analysis to account for multiple categories. Although inspired by intersectionality, this type of multi-axis disaggregated analysis fails to capture how systems of power and oppression give rise to qualitatively different experiences for individuals holding multiply marginalized identities [25].

Illustration While developing our CelebSET benchmark, we paid careful attention to balance it with respect to a crawled "ethnicity" label and binary gender. Although such a design is derived from the naturally occurring labels of the crawled and referenced datasets, the selected groupings have inherent limitations. Unlike the Pilot Parliaments Benchmark from the Gender Shades study [6] where the intersectional groups are defined with respect to skin type, "ethnicity" is an attribute that is highly correlated but not deterministically linked to racial categories, which are themselves nebulous social constructs, encompassing individuals with a wide range of phenotypic features [3]. Similarly, binary gender labels are compatible with the format of commercial product outputs, but exclusionary of those not presenting in the stereotypical representations of each selected gender identity [29].

We can see in Figure 1 that although CelebSET is balanced with respect to gender and coarse ethnicity labels, it is highly unbalanced with respect to age. At times, the exclusion of a particular group is unavoidable - for instance, our dataset lacks children because of the legal restrictions around the online exposure of a child. The average age of darker females in our benchmark is 37, with lighter males at 39, lighter females at 32, and darker males at 41. Since there are many more younger lighter females than older darker males, it is unclear whether the disparities between those groups is more correlated with age rather than race or gender. It is thus possible to optimize performance across gender and ethnicity categories in our benchmark, while continuing to perform poorly with respect to age (i.e. improve accuracy for only older darker males). This is an observation of the fairness gerrymandering effect [28] - where optimizing for fairness on one axis can compromise fairness in another.

3.2.3 Tension 3: Transparency and Overexposure. To limit misinterpretations of evaluation results on specific benchmarks, it is important to clearly communicate the limit of each benchmark and its appropriate context of use. Sharing details of the dataset development process with auditors and targets helps clarify the limit of the audit's scope, and the context in which results should be interpreted and appropriately acted upon. Similarly, publicly disclosing named audit targets can incite pressure to make the audit itself more impactful [39]. However, all this may come at great cost - such communications can also lead to targets overfitting to optimize product performance on the audit. Audits of this nature have also made institutions wary – in September 2019, IBM, an audit target in the Gender Shades study, removed its facial recognition capabilities from its publicly distributed API [26]. Similarly, Kairos began putting its services behind an expensive paywall following its inclusion in the Actionable Auditing study [27]. Such practices, although rightfully stopping developers from using a product revealed to be flawed, also compromise the product's auditability – making it more expensive and challenging for auditors to evaluate, even though it may still be in active use by enterprise customers.

Illustration In order to communicate the biases and limitations of CelebSET, we can create a datasheet [20] which helps clarify the context in which the benchmark should be used. This datasheet can specify, for instance, which demographic groups are covered in the audit, and what types of product applications the benchmark is best suited for. We can additionally note the small scope and limited demographic groupings of this particular audit. However, IMDB-Wiki, from which CelebSET and its variants are derived, is a publicly available online dataset [40]. This means the entire process of benchmark development is easily accessible to anyone including an audit target, who may decide to include the images in their training set. It is thus inevitable that the dataset will become obsolete as products overfit on the data.

4 RECONSIDERING THE ROLE OF ALGORITHMIC AUDITS

Our work shows that the algorithmic audit itself is a testing ground for the ethical concerns it is meant to evaluate. The audits need to be done with careful attention to the traps their targets fall into, and auditors must strive to live up to the ethical ideals they expect from their targets.

Auditors, thus, should approach these evaluations with a certain level of humility, acknowledging the limitations of their own evaluations and contextualizing each benchmark result as one component of a larger and more qualitative audit framework, which should begin by questioning the ethical use case of the product itself. The humble goal of the algorithmic audit is thus to expose blind spots rather than validate performance. Given its very own limitations and ethical concerns, FPT audits on benchmarks like CelebSET are a necessary but insufficient condition. They are inspections that can be used to stall or halt deployment, but do not have the weight of meaning to, by themselves, be used to justify FPT deployment or act as a condition for a moratorium. This means CelebSET as a benchmark should not be considered as a reward to game or a goal to strive for, but a very low bar not to be caught tripping over.

5 CONCLUSION

While designing CelebSET, an audit process for products employing facial processing technology (FPT), we were able to identify several ethical concerns with the developing norms of the algorithmic auditing of such products. These concerns in audit outcomes and processes often intersect with those of the audited product itself, as an unethical audit process can lead to a false sense of progress on the alignment of facial processing technology with the principles we have put forth. Both the audit process and the audited FPT, for instance, need to have careful privacy considerations, and avoid exploiting marginalized groups in the blind pursuit of increasing representation. If we take seriously the ethical expectations we have for the audited product, then we must also apply that same standard to the data and processes defining our evaluation.

REFERENCES

[1] Amazon. 2019. Amazon Rekognition FAQs. Retrieved Oct 31, 2019 from https://aws.amazon.com/rekognition/faqs/

[2] Ruha Benjamin. 2019. *Race after technology: Abolitionist tools for the new jim code.* John Wiley & Sons.

[3] Sebastian Benthall and Bruce D. Haynes. 2019. Racial Categories in Machine Learning. In *Proc. of the Conference on Fairness, Accountability, and Transparency (FAT).* 10.

[4] Sen. Roy Blunt. 2019. S.847 - Commercial Facial Recognition Privacy Act of 2019. https://www.congress.gov/bill/116th-congress/senate-bill/847/text

[5] Joy Buolamwini. 2018. When the Robot Doesn't See Dark Skin.

[6] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proc. of the Conference on Fairness, Accountability, and Transparency (FAT).*

[7] Matt Cagle and Nicole Ozer. 2018. Amazon Teams Up With Government to Deploy Dangerous New Facial Recognition Technology. (2018).

[8] Clarifai. 2019. Custom Face Recognition. Retrieved Oct 31, 2019 from https://www.clarifai.com/custom-face-recognition

[9] Rep. Yvette D. Clarke. 2019. H.R.2231 - Algorithmic Accountability Act of 2019. https://www.congress.gov/bill/116th-congress/house-bill/2231

[10] Rep. Yvette D. Clarke. 2019. H.R.4008 - No Biometric Barriers to Housing Act of 2019. https://www.congress.gov/bill/116th-congress/house-bill/4008/text?r=11&s=1

[11] Lord Clement-Jones. 2019. Automated Facial Recognition Technology (Moratorium and Review) Bill [HL] 2019-20. https://services.parliament.uk/bills/2019-20/automatedfacialrecognitiontechnologymoratoriumandreview.html

[12] Kimberle Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *The University of Chicago Legal Forum* (1989), 139.

[13] Kimberle Crenshaw. 2017. Kimberle Crenshaw on Intersectionality, More than Two Decades Later.

[14] Craig R Ehlen and Robert B Welker. 1996. Procedural fairness in the peer and quality review programs. *Auditing* 15, 1 (1996), 38.

[15] Norman L Enger and Paul William Howerton. 1980. *Computer Security: A Management Audit Approach*. Amacom New York.

[16] Face++. 2019. Face Attributes. Retrieved Oct 31, 2019 from https://www.faceplusplus.com/attributes/

[17] Sellywati Mohd Faizal, Mohd Rizal Palil, Ruhanita Maelah, and Rosiati Ramli. 2017. Perception on justice, trust and tax compliance behavior in Malaysia. *Kasetsart Journal of Social Sciences* 38, 3 (2017), 226–232.

[18] Sheera Frenkel. 2018. Microsoft Employees Question CEO Over Company's Contract With ICE. (2018).

[19] Timnit Gebru. 2019. Oxford Handbook on AI Ethics Book Chapter on Race and Gender. *arXiv preprint arXiv:1908.06165* (2019).

[20] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010* (2018).

[21] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. 2016. Ms-celeb-1m: Challenge of recognizing one million celebrities in the real world. *Electronic imaging* 2016, 11 (2016), 1–6.

[22] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. 2018. Gender recognition or gender reductionism?: The social implications of embedded gender recognition systems. In *Proc. of the ACM Conference on Human Factors in Computing Systems (CHI)*.

[23] Amy Hawkins. 2018. Beijing's Big Brother Tech Needs African Faces. Retrieved October 31, 2019 from https://foreignpolicy.com/2018/07/24/beijings-big-brother-tech-needs-african-faces/

[24] Michael Hind, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Alexandra Olteanu, and Kush R Varshney. 2018. Increasing Trust in AI Services through Supplier's Declarations of Conformity. *arXiv preprint arXiv:1808.07261* (2018).

[25] Anna Lauren Hoffmann. 2019. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society* 22, 7 (2019), 900–915.

[26] IBM. 2019. Release notes. Retrieved Oct 31, 2019 from https://cloud.ibm.com/docs/services/visual-recognition?topic=visual-recognition-release-notes

[27] Kairos. 2019. Kairos Face Recognition Pricing Guide. Retrieved Oct 31, 2019 from https://www.kairos.com/pricing

[28] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *Proc. of the International Conference on Machine Learning (ICML)*.

[29] Os Keyes. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proc. of the Human Computer Interact action* 2, CSCW, Article 88 (Nov. 2018).

[30] Steven Melendez. 2018. Despite a surge of tech activism, Clarifai plans to push further into government work. (2018).

[31] Steven Melendez. 2018. Uber driver troubles raise concerns about transgender face recognition. Retrieved October 31, 2019 from https://www.fastcompany.com/90216258/uber-face-recognition-tool-has-locked-out-some-transgender-drivers

[32] Michele Merler, Nalini Ratha, Rogerio S. Feris, and John R. Smith. 2019. Diversity in Faces. *arXiv preprints*, Article arXiv:1901.10436 (Jan. 2019), arXiv:1901.10436 pages.

[33] Microsoft. 2019. What is the Azure Face API? Retrieved Oct 31, 2019 from https://docs.microsoft.com/en-us/azure/cognitive-services/face/overview

[34] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2018. Model Cards for Model Reporting. *CoRR* abs/1810.03993 (2018). arXiv:1810.03993 http://arxiv.org/abs/1810.03993

[35] Paul Mozur. 2019. One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority. Retrieved October 31, 2019 from https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html

[36] Kristina Murphy. 2003. Procedural justice and tax compliance. *Australian Journal of Social Issues (Australian Council of Social Service)* 38, 3 (2003).

[37] Shruti Nagpal, Maneet Singh, Richa Singh, Mayank Vatsa, and Nalini Ratha. 2019. Deep Learning for Face Recognition: Pride or Prejudiced? *arXiv preprint arXiv:1904.01219* (2019).

[38] Mei Ngan, Mei Ngan, and Patrick Grother. 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.

[39] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Prof. of the Conference on Artificial Intelligence, Ethics, and Society*.

[40] Rasmus Rothe, Radu Timofte, and Luc Van Gool. 2016. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision (IJCV)* (7 2016).

[41] Silonie Sachdeva et al. 2009. Fitzpatrick skin typing: Applications in dermatology. *Indian Journal of Dermatology, Venereology, and Leprology* 75, 1 (2009), 93.

[42] Morgan Klaus Scheuerman, Jacob M Paul, and Jedr Brubaker. 2019. How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis and Image Labeling Services. (2019).

[43] Jacob Snow. 2018. Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots. Retrieved August 24, 2017 from https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28

[44] Olivia Solon. 2019. Facial recognition's 'dirty little secret': Millions of online photos scraped without consent. Retrieved October 31, 2019 from https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921

[45] Nisha Srinivas, Karl Ricanek, Dana Michalski, David S Bolme, and Michael King. 2019. Face Recognition Algorithm Bias: Performance Differences on Images of Children and Adults. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.