Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing

@NikoletaGlyn

Inioluwa Deborah Raji
University of Toronto

Timnit Gebru
Google Research

Margaret Mitchell
Google Research

Joy Buolamwini
MIT Media Lab.

Joonseok Lee
Google Research

Emily Denton
Google Research

Timnit Gebru
@timnitGebru

Apparently my manager's manager sent an email my direct reports saying she accepted my resignation. I hadn't resigned—I had asked for simple conditions first and said I would respond when I'm back from vacation. But I guess she decided for me :) that's the lawyer speak.

4:41 AM · Dec 3, 2020 · Twitter Web App

619 Retweets **371** Quote Tweets **5,438** Likes

FACIAL PROCESSING TECHOLOGY

a broad term that encompasses a variety of tasks ranging from face detection, facial analysis, and face verification or identification.

CelebSET

80 Celebrities

20 for each group

LM DM LF DF

Photo	Ethnicity	Colour	Gender	Age	Smile
		1 - 6	F/M	Photo Taken	Y/N

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%

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)**

Design Considerations

Consideration 1: Selecting Scope of Impact.

Consideration 2: Auditing for Procedural Fairness.

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%

Ethical Tensions

Tension 1: Privacy and Representation.

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%

Tension 2: Intersectionality and Group-Based Fairness.

Tension 3: Transparency and Overexposure.

@rajiinio
@timnitGebru
@jovialjoy
@mmitchell_ai
@cephaloponderer