United States House Committee on Science, Space and Technology

June 26, 2019

Hearing on

Artificial Intelligence: Societal and Ethical Implications

Written Testimony of

Joy Buolamwini
Founder, Algorithmic Justice League

Masters in Media Arts and Sciences, 2017, Massachusetts Institute of Technology
MSc Education (Learning & Technology), 2014, Distinction, University of Oxford
BS Computer Science, 2012, Highest Honors, Georgia Institute of Technology

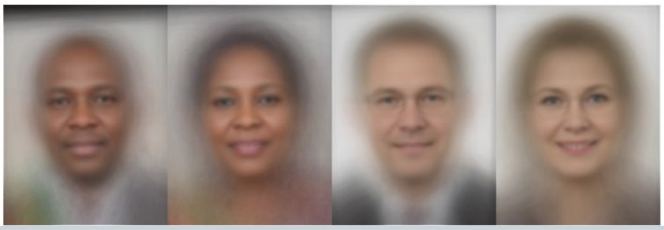
PhD *Pending*, MIT Media Lab

Made Possible By Critical Input from
Dr. Sasha Costanza-Chock
Injoluwa Deborah Raji

For additional information, please contact Joy Buolamwini at joy@ajlunited.org

Dear Chairwoman Johnson, Ranking Member Lucas, and Members of the Committee,

Thank you for the opportunity to testify on the societal and ethical implications of artificial intelligence (AI). My name is Joy Buolamwini, and I am the founder of the Algorithmic Justice League (AJL), based in Cambridge, Massachusetts. I established AJL to create a world with more ethical and inclusive technology after experiencing facial analysis software failing to detect my dark-skinned face until I put on a white mask. I've shared this experience of algorithmic bias in op-eds for Time Magazine and the New York Times as well as a TED featured talk with over 1 million views.[1] My MIT thesis and subsequent research studies uncovered substantial skin type and gender bias in AI services from companies like Microsoft, IBM, and Amazon.[2] This research has been covered in over 40 countries and has been featured in the mainstream media including FOX News, MSNBC, CNN, PBS, Bloomberg, Fortune, BBC, and even the Daily Show with Trevor Noah.[3]



Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Figure 1. Intersectional Skin Type and Gender Classification Accuracy Disparities. www.gendershades.org

---

[1] The Hidden Dangers of Facial Analysis, New York Times print run June 22, 2018, Page A25, online https://www.nytimes.com/2018/06/21/opinion/facial-analysis-technology-bias.html; Artificial Intelligence Has a Problem With Gender and Racial Bias. Here's How to Solve It, Time Magazine Optimist Edition http://time.com/5520558/artificial-intelligence-racial-gender-bias/; How I am Fighting Bias in Algorithms, https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms

[2] Joy Buolamwini, Timnit Gebru, Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification (February 2018), http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf; Inioluwa Raji, Joy Buolamwini, Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products (January 2019), http://www.aies-conference.com/wp-content/uploads/2019/01/AIES-19_paper_223.pdf

[3] See references of notable press mentions at www.poetofcode.com/press

Today, I speak to you as both a researcher and someone who has personally experienced algorithmic bias from flaws in AI systems and corporate hostility for publishing research showing gender and racial bias in an existing AI product.

In my testimony today, I will make 5 main points:

- First, the proliferation of AI in society across key social and economic areas makes it nearly impossible for individuals to avoid AI systems, and thus government and academia have an urgent responsibility to address the limitations of AI systems that can mask and further systematize structural inequalities.

- Second, harms from AI systems can arise from systems that propagate error (in)equity such that failures disproportionately impact select groups (i.e pedestrian tracking AI models failing more on children than adults) and processes that create a high exclusion overhead for individuals who fit outside of assumed norms (ie. trans* drivers being forced to undergo continuous and burdensome identification checks and ultimately denied economic opportunity).

- Third, the ability for AI systems to propagate sexism, racism, ableism, and ageism is documented and already marginalized groups like communities of color, low-income families, immigrants and people with disabilities are especially at risk for being further marginalized by AI systems used for employment, healthcare, government services, and policing.

- Fourth, sources of AI harms and bias can arise from lack of diversity in the field, misleading standard benchmarks, data collection and analysis processes, single-axis evaluation norms, and deprioritization of the public interest in the AI development, research, and education.

- Fifth and finally, government and academia must take actions to increase public awareness about the harms of AI, change academic and industry practices that obscure AI limitations, invest in diversifying the field, and ensure research on ethics, accountability, transparency and fairness in AI retains autonomy.

The Proliferation of AI in Society

We have arrived in the age of automation overconfident and underprepared. Often presented as a signifier of progress, artificial intelligence (AI) is increasingly influencing the lives of everyday people in ways that perpetuate individual and societal harms and can amplify past and present-day discrimination. Despite the danger that AI will entrench and exacerbate existing social inequalities, the promise of economic growth coupled with technological advances has spurred widespread adoption. In assessing the economic reach of AI, a recent Mckinsey report states “AI could potentially deliver additional economic output of around $13 trillion by 2030,

boosting global GDP by about 1.2 percent a year."[4] The public sector is also rapidly adopting AI to automate decision making, enhance judgement, improve civic engagement, and streamline interactions with common social services.[5] Taken together, the public and private sector embrace of AI makes it increasingly difficult to function in American society without encountering some form of this technology in consumer products or public services.

Even if an individual attempts to opt-out of an AI-fueled world, their neighbor may install a device with facial recognition enabled surveillance[6], or a bystander may upload a photograph of them to an online platform;[7] they may need to navigate streets increasingly populated with autonomous vehicles[8], submit a resume to an employer using undisclosed and unaccountable automated screening tools,[9] or otherwise interact with automated decision support systems that have already been shown by researchers to be biased and that violate privacy. As I will address more thoroughly already marginalized communities are often further marginalized by the use of these systems.

Select Examples of AI Harms

Though noble intentions like reducing fatalities and overcoming human biases animate the development of AI along with economic interests, research studies and headlines continue to remind us that AI applications are often imbued with bias that can lead to harms.

In identifying AI harms, we must pay particular attention to **error in(equity),** which arises when differential performance across demographics and phenotypic groups leads to harmful bias that disproportionately places the consequences of malfunctions on already marginalized or vulnerable populations (ie. purging of voter registration rolls that rely on automated name matching tools that are biased against non-traditionally European names results in limiting participation in democratic society)

---

[4]Bughin et al. "Notes from the AI Frontier : Modeling the Impact of AI on the World Economy", McKinsey Global Institute, (September 2018),

[5] "Essential Insights: Artificial Intelligence Unleashed", Accenture Federal Services, (2018), https://www.accenture.com/_acnmedia/PDF-86/Accenture-Essential-Insights-POV.pdf#zoom=50

[6] Rich Brown, "Nest says Hello with a new doorbell camera" (September 2017), https://www.cnet.com/news/nest-says-hello-with-a-new-doorbell-camera/

[7] Olivia Solon, "Facial recognition's 'dirty little secret': Millions of online photos scraped without consent" (March 2019), https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921

[8] Kirsten Korosec, "Uber reboots its self-driving car program" (December 2018) https://techcrunch.com/2018/12/20/uber-self-driving-car-testing-resumes-pittsburgh/

[9] Dipayan Ghosh, "AI is the future of hiring, but it's far from immune to bias" (October 2017) https://qz.com/work/1098954/ai-is-the-future-of-hiring-but-it-could-introduce-bias-if-were-not-careful/

We need to also attend to the **exclusion overhead** or the experiential differences that can emerge when technology forces certain demographic groups to expend more time, energy, and resources in an attempt to fit into systems that were optimised for a narrow group but used in a universal manner (ie. changing pitch of voice or speaking patterns to use voice recognition system).

INDIVIDUAL HARMS		COLLECTIVE SOCIAL HARMS
ILLEGAL DISCRIMINATION	UNFAIR PRACTICES	
HIRING		LOSS OF OPPORTUNITY
EMPLOYMENT		
INSURANCE & SOCIAL BENEFITS		
HOUSING		
EDUCATION		
CREDIT		ECONOMIC LOSS
DIFFERENTIAL PRICES OF GOODS		
LOSS OF LIBERTY		SOCIAL STIGMATIZATION
INCREASED SURVEILLANCE		
STEREOTYPE REINFORCEMENT		
DIGNITARY HARMS		

Table 1. Potential Harms from Automated Decision Making[10]

ERROR (IN)EQUITY

Transporting Risks: Which Lives Are We Saving?

According to the National Highway Traffic Safety Administration, vehicle fatalities killed an estimated 36,750 people last year in the United States,[11] and there is growing interest in the potential of autonomous vehicles to reduce deaths and increase transportation efficiency. Yet as Meredith Broussard reminds us in her book *Artificial Unintelligence,* the aspirational vision of what AI could potentially be is not an adequate substitute for reality.

Although autonomous vehicles have captured public and investor imagination, and companies like Tesla and Waymo are pushing the technology forward, development is in nascent stages Missteps including sensor driven fatalities, flawed system designs that enable external hijacking, and research showing pedestrian tracking can be less accurate in detecting dark-skinned

[10] See full chart: Lauren Smith, “Unfairness By Algorithm: Distilling the Harms of Automated Decision-Making” (December 2017) https://fpf.org/2017/12/11/unfairness-by-algorithm-distilling-the-harms-of-automated-decision-making/

[11]“Early Estimate of Motor Vehicle Traffic Fatalities in 2018”, US Department of Transportation (2018) https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812749

individuals, demonstrate the need for rigorous evaluations of AI-based vehicles that are entering public spaces.

Because autonomous vehicles must interface with humans, understanding the current performance and risks of the human-centered AI systems that inform car navigation (pedestrian tracking), safety features (drowsy driver alert) or passenger interactions (voice commands, biometric authentication) is critical to developing robust evaluation procedures. Furthermore, growing evidence, including the findings from my research on facial analysis systems, shows that human-centered AI products do not work equally well on different human populations. Differential performance across demographics and phenotypic groups can lead to harmful bias that disproportionately places the consequences of malfunctions on already marginalized or vulnerable populations.

At the recent workshop FATE at CVPR, a leading computer vision conference, an Oxford University researcher shared a study where they evaluated the accuracy of pedestrian detection algorithms. They found a statistically significant difference in the miss rate between adults and children across the top 24 performing algorithmic approaches.These findings along with the recent Georgia Tech study[12] shows that skin type influences the accuracy of state-of-the-art pedestrian tracking methods. These findings motivate concerns that autonomous vehicles that are positioned as lifesavers may in fact do the opposite. The Georgia Tech researchers attributed the difference in accuracy to the lack of representation of darker skinned individuals in training datasets used for pedestrian tracking. Training datasets for pedestrian tracking are not unique in having severe demographic skews. In addition, people with disabilities are seldom included in datasets for human-centered AI systems which further propagates ableism.

Thankfully, we are in the early days of AI development, and there is still time to course correct and exercise caution. Without robust evaluation methods to assess algorithmic vulnerabilities with AVs and high standards to evaluate the distribution of harms, keeping unproven technologies parked will preserve lives. When AI enabled technologies are presented as lifesavers, we must ask which lives will be saved? Which lives will matter?

EXCLUSION OVERHEAD

Hiring and Firing Bias: Who Looks the Part? Who Bears the Exclusion Overhead?

Unlike harmful practices explicitly linked to individual biases or systemic discrimination, AI systems are often perceived as being neutral,[13] making it even more challenging to identify and counteract machine-enhanced racism, sexism, ableism, and other harmful intersecting forms of discrimination. AI enabled tools are increasingly marketed as reducing human bias or being bias free. On the surface, this aim is laudable, but we must again separate potential from reality. The emerging use of AI to inform employment decisions demonstrates that even when AI builders

[12] Wilson et al. “Predictive Inequity in Object Detection” (2019) https://arxiv.org/abs/1902.11097
[13] Nicholas Carr , “The Glass Cage: Automation and Us” (2014)https://dl.acm.org/citation.cfm?id=2666139

hope to overcome human bias they may in fact mask the bias under the guise of machine neutrality.

On December 10, 2018, Upturn released a report detailing the integration of AI tools into human resources from screening to promotion and job termination.[14]Hiring intelligence company HireVue, one of the companies highlighted in the report, explicitly markets its products and services as reducing bias and increasing diversity. HireVue allows employers to interview potential job candidates on camera, by using AI to rate videos of each application according to verbal and nonverbal cues.[15] The system is reportedly trained on the current top performers of a company.[16] However, should exemplar employees be largely homogenous, there is a risk that the data-centric AI system learns to filter out applicants based on features protected by civil rights law (such as race or gender) rather than based on applicants' potential abilities to excel at the job. Amazon learned a similar lesson when an internal AI hiring tool developed to increase efficiency was reported to have harmful gender bias after the system was trained on ten years of hiring data. Unlike HireVue, Amazon's internal tool did not use video input - which introduces new additional risks - but was basing its discrimination on the inclusion of certain gendered descriptions. For instance, if the word "women's" and certain women's colleges appeared in candidates' resumes, the tool ranked them lower.[17]

As I wrote in a New York Times op-ed on June 22, 2018, "Given how susceptible facial analysis technology can be to gender and racial bias, companies using HireVue, if they hope to increase fairness, should check their systems to make sure it is not amplifying the biases that informed previous hiring decisions. It's possible companies using HireVue could someday face lawsuits charging that the program had a negative disparate impact on women and minority applicants, a violation of Title VII of the Civil Rights Act." The hope of overcoming bias cannot be a replacement for rigorous evaluations and external accountability. Beyond having companies implement internal bias mitigation processes, there needs to be external testing and validation to assess the use of AI in employment contexts, as well as regulatory oversight by knowledgeable agencies and consequences for those who violate civil rights law.

AI can serve not only as a gatekeeper for employment, but can also take on the role of terminator. For example, Uber has implemented automated authentication tools to verify that drivers on the platform are who they claim to be. The "Real Time ID Check" tool periodically notifies drivers to take images that are automatically compared to existing driver profile data.

---

[14] Miranda Bogen and Aaron Rieke, "Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias." (2018) https://www.upturn.org/reports/2018/hiring-algorithms.

[15] Corporate Financial Institute, "HireVue Interview Guide: How to prepare for a HireVue interview," accessed on 20 May 2019 https://corporatefinanceinstitute.com/resources/careers/interviews/about-hirevue-interview/

[16] https://www.businessinsider.com/hirevue-ai-powered-job-interview-platform-2017-8

[17] Jeffrey Dastin "Amazon scraps secret AI recruiting tool that showed bias against women" (October 2018) https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

However, the system has limitations. On May 20 2019, Mr Willam Fambroug sued Uber for $227,033 in reparations and punitive damages the company automatically deactivated his account with no means to contest the situation.[18] He states in his legal filing:

> "Uber uses face recognition to verify the correct driver is using the correct driver account. It is universally known, face recognition apps have problems recognizing the "Black" skin color… When asked to verify, .. the app does not recognize my selfie. Uber favors whites who have no problem with the app over blacks who do, as shown by the reasons Uber states for my account deactivation."

This is not an isolated incident nor one that extends only to skin pigmentation. Multiple transgender Uber drivers reported that the feature repeatedly locked them out.[19] Uber reportedly deactivated the accounts of transgender drivers,[20] erroneously denying economic opportunity and highlighting how trans* and gender non-conforming people face additional harms from AI-based tools that are not designed to accommodate a broad range of gender identities and expressions.[21] One former transgender driver with a deactivated account shared that over an 18 month period the Uber system requested over 100 checks for account validations, and that they were suspended from the app for photo inconsistencies as a result. We have to keep in mind not just discriminatory outcomes of AI tools but also the experiences of those using these systems.[22]

These checks require that the driver pull over to take the photo, limiting productivity and time to earn money. I use the term the "exclusion overhead" to capture the experiential differences that can emerge when technology forces certain demographic groups to expend more time, energy, and resources in attempting to fit into systems that were optimised for a narrow group but used in a universal manner. Designers and researchers of AI systems must attend to the exclusion overhead and also keep in mind that AI tools can mask and systematize harmful discrimination.

The use of AI in transportation and employment demonstrate just a handful of ways well intentioned AI tools can propagate harms. Table 1. highlights some additional areas where AI

---

[18] "William Fambrough Vs Uber Technology Inc." May 20 2019 civil suit. Details here: https://drive.google.com/open?id=0B_IVlfmguHPNSFczNVNfWUNzYnNHN21OYVFiZXN2dmdSME9F

[19] https://www.them.us/story/trans-drivers-locked-out-of-uber

[20] Jaden Urbi, "Some transgender drivers are being kicked off Uber's app" in CNBC (August 2018) https://www.cnbc.com/2018/08/08/transgender-uber-driver-suspended-tech-oversight-facial-recognition.html

[21] See more about the harms trans* and gender non-conforming people face from automated decision making systems: Sasha Costanza-Chock, "Design Justice, A.I., and Escape from the Matrix of Domination" in Journal of Design and Science (July 2018), https://jods.mitpress.mit.edu/pub/costanza-chock . For the limitations and harms of binary gender classification see: Os Keyes,"The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition" (2018), https://ironholds.org/resources/papers/agr_paper.pdf

[22] Jaden Urbi, "Some transgender drivers are being kicked off Uber's app" in CNBC (August 2018) https://www.cnbc.com/2018/08/08/transgender-uber-driver-suspended-tech-oversight-facial-recognition.html

systems can limit access to opportunity, render undue economic loss, and perpetuate social stigma. Some areas highlighted in the chart like housing, employment, education, and credit lending have federal protections which make it paramount that we develop AI in a manner that doesn't undercut existing protections and that we educate researchers and practitioners on existing laws. Books like Mireille Hildebrandt's "Law for Computer Scientists and Other Folk" offer a primer to help educate computer scientists on legal matters as the scope of their creations impact society writ large.[23]

Other areas that can lead to collective social harms such as stereotype reinforcement and increased surveillance require an increased awareness of how historic inequalities and controlling narratives shape seemingly objective technologies. In her award-winning book *Dark Matters*, Simone Brown underscores how historic and ongoing oppression - particularly antiblackness - shapes present-day surveillance technologies. And as Shoshana Zuboff emphasizes in her book *Surveillance Capitalism*, the data gathering that fuels large technology companies and recent advancement in AI perpetuate power asymmetries in such a manner where participating in everyday life necessitates submitting to invasive tracking. Both Brown and Zuboff offers insights that can help AI practitioners and researchers better understand how the identification, classification, and measuring of individuals can be used for social control and to deepen entrenched inequalities.

The ExCoded: Further Marginalizing the Already Marginalized

AI Systems Reflect Society

Ultimately, society shapes technology and the shape of American society is one which was built on the genocide and displacement of Indigenous peoples; slavery; the oppression of communities of color, one that did not give women full standing as citizens until the 20th century and still contends with gender discrimination, one that propogated scientific racism, one with a technology industry that is prodimantly led by white men, and one that has allowed corporate interests to influence the policy makers meant to advance the public interest. As such, we have a situation where a small largely homogenous group of people are designing the AI technologies that increasingly touch all of our lives. Without interventions that look at how social and historical factors shape AI development, research, and education, we will increase the technical capabilities of AI systems in ways that continue to worsen inequalities.

For example, AI used to determine hiring decisions has been shown to amplify existing gender discrimination. Law enforcement agencies are rapidly adopting predictive policing and risk assessment technologies that have been shown to reinforce patterns of unjust racial discrimination in the criminal justice system[24]. AI systems also determine the information we see

---

[23] "Law for Computer Scientists" https://lawforcomputerscientists.pubpub.org/

[24] Kristian Lum, William Isaac. "To predict and serve?" (October 2016), https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1740-9713.2016.00960.x ; Rashida Richardson et al. "Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice" (March 2019), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3333423

on social media feeds, and can perpetuate misinformation, amplify hate speech, and unwittingly promote the sexualization of very young children[25] when optimized to prioritize attention-grabbing content.[26] In a world where AI systems influence access to opportunity, freedom, and information, we must attend carefully to the risks they pose and to the distribution of benefits and burdens they produce.

How AI Stigmatizes Cultural Signifiers and Online Behavior

In particular, the burdens of AI fall disproportionately on populations that have been historically excluded from exercising power and obtaining full rights due to patriarchy, white supremacy, and other intersecting forms of oppression. For example, studies of natural language processing (NLP) models that are increasingly used to analyze text for sentiment have revealed how these models often reinforces stereotypes[27],negative associations[28], and misunderstandings of culture.[29] Furthermore, the vast majority of NLP models are trained on what is deemed standard English, making these systems especially ill-equipped to deal with patterns of language such as patois or cultural variations that are not legitimated by state power. Despite these issues, government agencies have explored the use of social media content analysis for extreme vetting to determine who is deemed acceptable and who is deemed a threat[30].

Being labeled suspicious either because your patterns of behavior fit outside what has been defined as normal by an AI system inheriting the power norms of a society, because you belong to a stigmatized group, or because you refuse to submit your activities to algorithmic evaluation can impinge opportunities. In a landmark study on algorithmic bias Dr. Latanya Sweeney, the former chief technologist of the FTC, demonstrated that online searches for names coded as African-American were more likely to bring up search ads associated with arrest records regardless of whether or not the individual actually had a record. In doing due diligence, an employer, landlord, or social worker who searches a stigmatized name may be more likely to dismiss an individual simply because of the risk implied by a negatively associated ad.

Moreover, due diligence is now being automated by AI tools. One company, Predictim, provides a service to conduct background checks on babysitters in part by performing a social media analysis to determine risk ratings for bullying, harassment, being disrespectful and having a bad attitude" Parents are notified whether or not a prospective candidate submits to the search, and

---

[25] Max Fisher and Amanda Taub "On YouTube's Digital Playground, an Open Gate for Pedophiles" (June 2019), https://www.nytimes.com/2019/06/03/world/americas/youtube-pedophiles.html

[26]The Spread of True and False News Online: http://science.sciencemag.org/content/359/6380/1146

[27] Bolukbasi et al. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings" (June 2016), https://arxiv.org/abs/1607.06520

[28] Caliskan et al."Semantics derived automatically from language corpora contain human-like biases" (April 2017) https://science.sciencemag.org/content/356/6334/183.abstract

[29] Su Lin Blodge and Brendan O'Connor, "Racial Disparity in Natural Language Processing:A Case Study of Social Media African-American English" (June 2017),https://arxiv.org/pdf/1707.00061.pdf

[30] Aaron Cantú and George Joseph, "Trump's Border Security May Search Your Social Media by 'Tone'" (August 2017), https://www.thenation.com/article/trumps-border-security-may-search-your-social-media-by-tone/

failure to provide access to personal social media accounts can raise suspicion. Malissa Nielsen, a 24-year-old babysitter who stated she had nothing to hide, submitted her social media information and was surprised to find she was flagged, losing her job in the process. The company does not reveal how these determinations are made, despite the impact they can have over life altering decisions on employment. When AI tools attempt to reduce complex language or behaviour patterns to make unsubstantiated inferences about a person or perpetuate cultural stigma, individuals who belong to communities that have been othered and criminalized will suffer most.

AI Risks for Immigrants, Muslims, and Low-Income Families

Furthermore, those who are in vulnerable situations like refugees seeking asylum or those who face large power asymmetries like immigrants seeking visas, are under increased pressure to subject themselves to algorithmic evaluation or be labeled suspicious for daring to protect their privacy or assert their dignity. For example, recently, the Department of Homeland Security began requiring all visa applicants (15 million people per year) to submit email and social media account information, despite USCIS internal evaluations that show the failure of algorithmic analysis of social media to identify risky actors, and over widespread objections about the potential misuse and harms of automated analysis and classification of immigrants based on social media information. The Brenna Center found Muslims are particularly vulnerable to targeting.[31]

Class dynamics also influence the distribution of burdens from AI systems. In her book *Automating Inequality,* Political Scientist Virginia Eubanks highlights how low-income communities have long been used as guinea pigs to test automated decision making tools. She also illuminates how race and class work together to deepen existing inequalities when top-down tools are introduced into social work. Eubanks chronicles the development of a system implemented in the call screening center for the Allegheny County Office of Children, Youth and Families (CYF) child neglect and abuse hotline to to forecast child abuse and neglect, called the Allegheny Family Screening Tool (AFST)[32]. However, the tool is heavily biased towards predicting children in families with the least resources as being abused, and often overlooking serious cases of neglect in more resourced households. As a result, parents from lower income households get more frequently flagged by the CYF, even despite less evidence of child maltreatment, and are thus at greater risk of losing custody of their children.

As a study on algorithmic risk assessment in child services describes, such cases, although involving a complex analysis of pros and cons, are more likely to become detrimental to low income families[33]:

---

[31] For an in-depth overview of DHS monitoring of social media, see the Brennan Center report at https://www.brennancenter.org/publication/social-media-monitoring; For objections to this kind of monitoring, see http://bit.ly/dhs-social-comments-bu.

[32] Virginia Eubanks,"A Child Abuse Prediction Model Fails Poor Families" (January 2018), https://www.wired.com/story/excerpt-from-automating-inequality/

[33] Chouldechova et al. "A case study of algorithm-assisted decision making in child

> "There is a possibility that some communities— such as those in poverty or from particular racial and ethnic groups—will be disadvantaged by the reliance on government administrative data."

A similar phenomenon of undue burden is observed the allocations of social services. In 2003, a California court case ruled in favour of a welfare system requiring the use of fingerprint identification for aid recipients. The system was implemented as a measure against fraudulent or multiple applications for government aid. However, the prosecuting party also claimed that such measures minimized the impact of the program, deterring immigrants and those experiencing poverty, who were more likely uncomfortable with the practice, from participating and receiving the aid they needed. She also presented the case of these vulnerable groups being justified in their discomfort, as the mandatory fingerprinting, exposing their identifiable biometric data, posed a threat to their overall privacy and safety.[34]

How Automated Proxies Amplify Racism in Price Discrimination and Policing

In reflecting societal patterns, designers of AI models have found the use of zip code to be a powerful variable for making inferences about individuals, because where you live can convey critical information about your place in society including socioeconomic factors like income, education, and employment. The use of zip code can also serve as a proxy that makes decision making seem more neutral by obscuring how geographic locations map to demographics, historic oppression, and ongoing inequalities. The obscuring naturing of zip code coupled with its correlation to demographic factors like race have made it an ideal variable to provide a veneer of objectivity,

In the past, the use of zip code has been intentionally employed to limit material resources and opportunity to the already privileged. The practice of redlining has a long legacy in the United States. Building on patterns of racial segregation, redlining was historically used to keep racial minorities in their place. By categorizing specific zip codes as off limits for receiving loans, raising rates on insurance for minority neighborhoods, and gatekeeping particular neighborhoods to dissuade racial integration, the practice of redlining codified by the passage of the National Housing Act of 1934,[35] was explicitly deployed to preserve opportunities for white communities. Today, AI systems that incorporate geolocation data can learn patterns of exclusion and exploitation. Elevating the visibility of the Tiger Mom Tax, a 2015 study found that test preparation services customers in zip codes with a high density of asian residents were being charged twice the price for services as compared to the average price of these services[36].

---

maltreatment hotline screening decisions" (2018), http://proceedings.mlr.press/v81/chouldechova18a/chouldechova18a.pdf

[34] "Sheyko v. Saenz" civil suit details here: https://cite.case.law/cal-app-4th/112/675

[35] Kevin Fox Gotham, "Racialization and the State: The Housing Act of 1934 and the Creation of the Federal Housing Administration" (2000), https://journals.sagepub.com/doi/10.2307/1389798

[36] Vafa et al. "Price Discrimination in The Princeton Review's Online SAT Tutoring Service"(September 2015), https://techscience.org/a/2015090102

Zip codes are also used in predictive policing applications to indicate areas to patrol for crimes. However, the information that is used is based on past information about areas that have been policed. Given that black and brown neighborhoods are overpoliced and crime in other location is not recorded to the same extent, what might on the surface seems like an objective tool for law enforcement instead reinforces the status quo while using AI to legitimize racialized policing practices.

Colorism, Ageism, and Ablesim in AI for Healthcare

In addition to harms that can arise when AI tools learn or reinforce patterns of discimination, these tools can also lead to bad outcomes when differences between individuals are ignored or erased. Making one group the standard, by which all others must fit can counteract the very benefits designers of AI systems hope to achieve. Studies that highlight breakthroughs in AI for specific domains at times use language that suggests universal progress, when the reality shows a different story. In 2017 Stanford University researched released a study announcing a technical breakthrough in assessing melanoma.[37] The AI systems developed by the researchers matched the accuracy rates comparable to that of dermatologists. However the dataset used for evaluation was overwhelmingly composed of lighter-skinned individuals, even though people with darker skin can get skin cancer.[38] If this model were to be commercialized and used to assess individuals for melanoma, people with skin variations not included in the dataset might have serious problems that could be missed. Already individuals with darker skin are less likely to be diagnosed with melanoma until more advanced stages.[39] Building inclusive AI-enabled diagnostics could help reverse this trend, but only if researchers in the field are intentional.

Age and ability are factors that can influence the technical performance of AI systems built for healthcare. In a February 2019 study, researchers demonstrated the existence of algorithmic bias in state-of-the-art facial expression and landmark recognition methods, which affects the performance of these algorithms for older adults with cognitive impairment.[40] Used as is, the algorithms were less accurate on older adults before being specifically adapted to the population showing that when issues are detected mitigation strategies may be employed. However, they found that even when they attempted to train the algorithms to work better on the faces of older adults with dementia, there were still statistically significant differences between older adults with dementia and those without. The study indicates that not all clinical populations will have the same accuracy even when state-of-the art algorithms are applied to in a wide range of potential health applications including clinical assessment of depression, detection of pain in non-communicative individuals, monitoring progression of motor neuron disease, and alternative interfaces for differently abled persons.

---

[37] Esteva et al. “Dermatologist-level classification of skin cancer with deep neural networks” (February 2017), https://www.nature.com/articles/nature21056

[38] Porcia T. Bradford, “Skin Cancer in Skin of Color” (August 2009), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2757062/

[39] Same as above.

[40] Babak Taati et al., “Algorithmic Bias in Clinical Populations — Evaluating and Improving Facial Analysis Technology in Older Adults With Dementia” (February 2019) https://ieeexplore.ieee.org/document/8643365

As the examples above indicate, we cannot afford to assume AI tools will be bias-free or harmless precisely because these tools, when used in the real world, are part of societal processes that have been shaped by racism, sexism, ableism, and other harmful forms of intersecting discrimination.

Addressing Algorithmic Harms and Bias

Since AI systems are influencing all sectors of society and have documented harms that can increase inequality and facilitate mass surveillance, we must design the processes that shape AI development, research, and education to anticipate, identify, mitigate, and redress harms. Organizations like AI Now and Data & Society have conducted extensive studies that demonstrate the need for interdisciplinary research and policy work that take into account the social and historical contexts that shape the design, development, and governance of AI systems.[41] Tools like algorithmic impact assessments and processes for thinking through legislating AI are crucial components for addressing the harms of AI that extend beyond narrow technical solutions. In addition, AI researchers are starting to contend with issues of ethics, fairness, transparency, and accountability, and there are now a growing number of workshops and conferences in this expanding area of research including the Fairness, Accountability and Transparency (FAT*) conference as well as the AI Ethics and Society (AIES) conference. These combined efforts have helped to spotlight societal sources of AI harms and bias as well as highlight failings in the research and development of AI that have masked problems.

If we are not intentional about designing AI systems with equity in mind, we will replicate existing structural inequalities. With this in mind, below I outline some areas of concern with the state of AI in the United States and their implications for AI harms and bias below. I follow up with my personal experience as an algorithmic bias researcher from a severely underrepresented group (black women) in the domain of AI to provide real-world context to these issues I've encountered firsthand.

- **PRIVILEGED IGNORANCE:** The vast majority of researchers, practitioners, and educators in the field are shielded or removed from the harm that can result in the use of AI systems leading to undervaluation, deprioritization, and ignorance of problems along with decontextualized solutions. The communities most likely to be harmed by AI systems are least likely to be involved in the teaching, design, development, deployment, and governance of AI; even when underrepresented individuals enter previously inaccessible spaces, we face existing practices, norms, and standards that require

[41] AI Now Reports https://ainowinstitute.org/reports.html; Jessie Daniels et al., "Advancing Racial Literacy in Tech" (May 2019) https://datasociety.net/wp-content/uploads/2019/05/Racial_Literacy_Tech_Final_0522.pdf;Kadija Ferryman and Mikaela Pitcan, "Fairness in Precision Medicine" (February 2018) https://datasociety.net/wp-content/uploads/2018/02/Data.Society.Fairness.In_.Precision.Medicine.Feb2018.FINAL-2.26.18.pdf

system-wide not just individual change (ie. Well-meaning organization builds tool to automate screening of children who may be abused or neglected only to make it more likely children who are not at risk but come from low-income families to be targeted)[42]

- **MISLEADING EVALUATIONS:** The field suffers from a false sense of universal progress in part due to misleading evaluation norms and industry wide datasets with significant demographic, phenotypic, and geographic skews. The current push for AI fairness research risks establishing new computational approaches that mask societal problems which cannot be addressed through isolated technical solutions (ie. researchers and practitioners evaluate AI performance based on biased gold standard benchmarks.)
- **PROBLEMATIC DATA FLOWS:** A common response to uncovering severe data imbalances is to collect more data; however, how data is collected, categorized, and distributed presents ethical challenges around consent and privacy along with societal challenges with the politics of classifications where social categories like race and gender become reified into the technical systems that increasingly shape society.
- **SINGLE-AXIS ANALYSIS:** Emerging algorithmic bias research tends to focus on a single-axis of discrimination like race or gender in isolation, missing risks for populations who encounter intersecting forms of discrimination like women of color who contend with both racism and sexism working in combination. Without an intersectional lens our understanding on the scope, spread, and impact of AI harms and bias will be limited. (ie. Government funded human-centered AI research fails to require intersectional analysis echoing issues of government funded health studies in the past not requiring clinical studies data to be disaggregated.)
- **EROSION OF PUBLIC INTEREST:** The risks associated with AI harms and bias threaten trust in government agencies as well as the reputation and product acceptability of influential technology companies who are increasingly funding research and influencing policy discussion around ethics, transparency, accountability, and fairness in AI. Without explicit measures to address conflicts of interest and to protect researchers whose work for the public interest are in tension with private interests, public-private partnerships can lose legitimacy and critical AI harms, research may be silenced, sidelined, and/or underfunded. (ie. Amazon sponsors NSF AI Fairness research despite public company hostility to AI Fairness researchers.)

[42] Virginia Eubanks, “Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor” (2018)

PRIVILEGED IGNORANCE

My experiences as one of few black women working on algorithmic bias research has shown me firsthand the importance of having people who are impacted by AI harms and bias working in the field. During the 2015-2016 school year as a masters student at MIT, I had the experience of putting on a white mask in order to have my dark-skinned face consistently detected by face tracking software I incorporated into a coding project. The system I built worked fine on my lighter-skinned colleagues. Like many practitioners, I adopted the practice of using preexisting code made available on the internet in order to integrate the face tracking features. Like many code packages that contain AI models, there was no indication that the system might work better on some groups than others. Without having access to the training data or details about the underlying AI model, I was operating in the dark. This experience of coding in white face motivated the research that became my master's thesis. For this work, I evaluated facial analysis systems from leading tech companies including IBM and Microsoft on the task of guessing the reductive binary gender of a face. All companies performed better on lighter faces than darker faces, and all performed better on male-indentified faces than female-identified faces. When I did an intersectional analysis looking at gender and skin type in combination, I found that error rates were no more than 1% for lighter-skinned men but they soared to over 30% for darker skin women in the worst case. The 2018 research paper published from these findings was widely publicized and covered by national and international media. The public attention lead to private sector action with IBM, Microsoft, and other companies operating in the face space referencing the work in relation to developments on their facial analysis services. The attention also gave me the opportunity to speak to practitioners and researchers inside various companies not just the ones I initially audited. I heard one of three stories:

1. **Decision Makers Deprioritize Issues They Deem Irrelevant or Inconvenient:** In some organizations, junior members of teams reported seeing indications of trouble but senior leadership failed to prioritize issues around algorithmic bias. Among the most troubling case I encountered was from the lead of quality assurance for a company, who expressed regret at not testing accuracy on darker-skinned faces because it would have required more effort than deemed necessary to sell the product.
2. **Homogenous Teams Lack Diverse Perspectives:** In other organizations, despite having access to competitive talent pools, the teams were not aware of the extent of demographic and/or phenotypic bias of their face products and had not explored considering skin type as a variable for analysis
3. **Diverse Leadership Does Not Provide Immunity:** In organizations with people of color in executive roles, I learned that some were aware of bias issues through personal experience and were working to counteract the issues. Still, their products remained biased in part because existing state-of-the-art models and readily available data were biased.

In short, market pressures combined with priorities and constraints of those with the power to create the types of products I scrutinized contributed to the selling of biased AI products. The first two kinds of stories emphasize the importance of having more diverse decision makers, researchers, and practitioners when it comes to identifying, prioritizing, and to some extent empathizing with an issue. The third scenario shows that diversity, while necessary in surfacing issues, is a starting point. Counteracting bias in AI requires not just more inclusive hiring practices.It also requires a multi-pronged approach to shift norms, standards, and incentives, as well as to ensure meaningful external oversight, pressure, regulatory intervention, procurement policies, and significant redress mechanisms for communities that are harmed by biased AI.

MISLEADING EVALUATIONS

In addition to diversifying the practitioners, the practices that shape a field particularly those that have been largely homogenous like AI must also be changed. A pressing question I had as I conducted research which uncovered some of the largest recorded gender and skin-type accuracy disparities in commercial facial analysis was, “Despite my own experiences with technical failures with facial analysis technology, why was I continuing to hear about universal breakthrough about research in the area?” In reviewing key papers on advances in the facial analysis research, I found evaluations of performance generally centered on de facto industry benchmarks. **However, the benchmarks that are used to evaluate the performance of AI systems often have significant representational limitations. Impressive performance on a gold standard can indicate advancement, but if the gold standard only includes data from pale males, we have to ask - improvement for who?**

In 2014, Facebook researchers published the landmark Deep Face paper. Using deep learning techniques, they made a significant leap on the performance of the gold standard facial recognition dataset of the time Labeled faces in the Wild (LFW). They recorded a 97.35% accuracy on LFW significantly exceeding the prior top performance.[43] This was a widely recognized breakthrough and demonstrated the effectiveness of using deep neural networks for computer vision. However, work exploring the demographic composition of LFW, found that the benchmark was 77.5% male and 83.5% White.[44] These overwhelming demographic imbalances persist in core datasets across different domains of AI and limit our understanding of the performance of models on populations that are either severely underrepresented or excluded from benchmarks datasets. The table below provides information about notable imbalances by age, gender, and/or skin type for seven prominent face datasets.

---

[43] Yaniv Taigman et al., “DeepFace: Closing the Gap to Human-Level Performance in Face Verification” (June 2014) https://www.cs.toronto.edu/~ranzato/publications/taigman_cvpr14.pdf

[44] Hu Han, Anil K. Jain, “Age, Gender and Race Estimation from Unconstrained Face Images”(July 2014) http://biometrics.cse.msu.edu/Publications/Face/HanJain_UnconstrainedAgeGenderRaceEstimation_MSUTechReport2014.pdf

	Age Group							Binary Gender[45]		Skin Color / Type	
Dataset	**0-3**	**4-12**	**13-19**	**20-30**	**31-45**	**46-60**	**>60**	**Female**	**Male**	**Darker**	**Lighter**
LFW	1.0%	10.6%	25.4%		29.6%		33.4%	22.5%	77.4%	18.8%	81.2%
IJB-C*	**0.0%**	**0.0%**	**0.5%**	**16.2%**	**35.5%**	**35.1%**	**12.7%**	**37.4%**	**62.7%**	**18.0%**	**82.0%**
Pub fig	1.0%	10.8%	55.5%		21.0%		11.7%	50.8%	49.2%	18.0%	82.0%
CelebA	77.8%					22.1%		58.1%	42.0%	14.2%	85.8%
UTKface	8.8%	6.5%	5.0%	33.6%	22.6%	13.4%	10.1%	47.8%	52.2%	35.6%	64.4%
AgeDB	0.1%	0.52%	2.7%	17.5%	31.8%	24.5%	22.9%	40.6%	59.5%	5.4%	94.6%
IMDB-Face	0.9%	3.5%	33.2%	36.5%	18.8%	5.4%	1.7%	45.0%	55.0%	12.0%	88.0%

Table 2. Age, Binary Gender, and Skin Color/Type Distribution of 7 Prominent Face Datasets
Data reproduced from IBM Research Diversity in Faces Report: https://arxiv.org/pdf/1901.10436.pdf
*IJB-C is a US Government Face Dataset
Produced by the National Institute for Standards and Technology

Moving forward, the field must examine the appropriateness of the metrics and benchmarks by which we measure success and make it common practice to disclose the demographic composition of evaluation benchmark datasets to better assess which populations are either underrepresented or excluded. This basic step for transparency will provide a more realistic view of technical progress.

PROBLEMATIC DATA FLOWS

A common response to uncovering severe data imbalances is to collect more data; however, how data is collected, categorized, and distributed presents ethical challenges around consent, privacy, and compensation along with societal challenges with the politics of classifications. Further, the use of data can provide a veneer of neutrality and objectivity that belie the subjective choices made in selecting and analyzing data.Yes, the rapid adoption of AI in recent years has been made possible by the data surge and increased computation power of the 21st century that now fuels machine learning techniques developed in the 20th century. Machine learning has become the ascendant approach to AI, as the gathering of immense data enables people to use learning algorithms to build models aimed at tasks ranging from classifying faces to identifying disease. **For data-centric technology like machine learning enabled AI, data is destiny. Yet the data that is fueling AI is not neutral. For example, when machines learn from historic practices, they can reinforce past inequalities instead of overcoming them.** Sexist hiring managers or discriminatory recruitment methods are replaced by faceless AI tools that unfairly deny economic opportunity with data-driven precision.

[45] No systematic information is yet available about face based biometric identification system failure rates for gender nonconforming, nonbinary gender, agender, and/or transgender people, specifically.

The flow of data in common AI development pipelines introduces bias at multiple points. Given a generalized overview of a machine learning model development pipeline, there are several areas where bias can be introduced along the way. Understanding how data flows in the practice of building AI models can help with identifying points of intervention, but we must also interrogate how the data that enters a pipeline is obtained.

Current data harvesting processes eschew consent and violate expectations of privacy. Returning to the face space, we see that often convenience sampling is used. Given the availability of online images, researchers and companies scrape the internet for photos generally collected without consent and in violation of expectations of privacy, as Adam Harvey demonstrates in the MegaPixel project. Even when consent is given to store data in one context, scope creep can make it tempting and all too easy for companies storing personal photos to use those images for another purpose.[46]

Classification schema can reify social constructs and limit analysis. The categorization of data with labels to feed into various AI pipelines often relies on existing classification taxonomies for factors like race, which are socially constructed. While these labels can be useful for conducting disparity audits, they can also risk reifying certain categories and limit analysis. For example, the concept of race, which changes over time, and geography does not denote specific stable physical characteristics and there can be significant intraclass variation. As such using race as a category for evaluating human-centered computer vision tasks can yield poor results compared to use of phenotypic characteristics like skin type, which does not exclusively belong to one socially constructed racial group.[47] Similarly, the common use of binary (Male/Female) gender classification in AI systems may systematically erase the existence of trans*, nonbinary, and gender-nonconforming people, with real-world discriminatory impacts.[48]

The labor and labeling practices used to process data can perpetuate inequality. Adding to the challenges of choosing classification schema, the application of labels from that schema is often facilitated by employing human annotators who introduce their own individual bias into a labelling process, may be unaware of how their efforts are being utilized (as was the case of worker providing labels to power computer vision applications intended for military

---

[46] James Vincent, “A photo storage app used customers’ private snaps to train facial recognition AI” in The Verge (May 2019) https://www.theverge.com/2019/5/10/18564043/photo-storage-app-ever-facial-recognition-secretly-trained

[47] Cynthia M. Cook et al., Demographic Effects in Facial Recognition and Their Dependence on Image Acquisition: An Evaluation of Eleven Commercial Systems (February 2019)

[48] Os Keyes, “The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition” (November 2018) https://ironholds.org/resources/papers/agr_paper.pdf; Heath Fogg Davis, “Beyond Trans: Does Gender Matter” (September 2018) https://nyupress.org/9781479855407/

operation), and are poorly paid for their effort[49]. Furthermore human annotators and can introduce additional personal bias making it of particular importance to make sure we properly document data sources, labeling processes, and classification scheme limitations.

As data protections vary in different regions, those with the least protections risk the greatest exploitation.

Beyond providing comprehensive documentation of data collection using approaches like Datasheets for Datasets or Data Nutrition Labels,[50] we must also examine currently accepted research and development practices for mass scale data collection that eschew consent and can violate privacy particularly as legislation like GDPR extends protection to digital information for EU citizens. As data protections vary across jurisdiction we must also be aware of dynamics that can amplify AI and Data Colonialism where individuals with the least protection -in particular, those from the Global South - are the most exploited.[51]

SINGLE-AXIS ANALYSIS

In aiming to address issues of AI harms and bias, an interdisciplinary approach is necessary to provide critical social and historical context as AI is applied in various domains. We need to also make sure that research and evaluation mechanisms like technical standards being developed are adapting to incorporate insights from social sciences. In 1989, legal scholar Kimberlé Crenshaw demonstrated that single-axis antidiscrimination protections by race or by gender were insufficient to protect multiply-burdened groups (in particular, Black women) in the courts. She showed courts repeatedly rejected Black women's discrimination claims when they could only prove that they had been discriminated against specifically as Black women - in other words, their claims were not legally actionable unless they could statistically prove that firms were discriminating either against all women, or against all Black people. Building on Crenshaw's insight, a major focus of my algorithmic bias research has been championing the relevance of intersectional analysis in the domain of human-centered AI systems.

As AI systems are being used for cases like law enforcement, housing, or employment, they must be externally evaluated to assess suitability of use on intended populations should there be legislative approval for deployment. Such evaluations cannot rely on a single aggregate metric for accuracy and must be constructed to disaggregate differences between subpopulations, which can be substantial.

---

[49] Mary L. Gray, Siddharth Suri, “Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass” (2019)

[50] Timnit Gebru et al., “Datasheets for Datasets” (April 2019) https://arxiv.org/pdf/1803.09010.pdf

[51] Amy Hawkins,”Beijing’s Big Brother Tech Needs African Faces” in Foreign Policy (July 2018) https://foreignpolicy.com/2018/07/24/beijings-big-brother-tech-needs-african-faces/

GENDER CLASSIFIER	TYPE I	TYPE II	TYPE III	TYPE IV	TYPE V	TYPE VI
Microsoft	1.7%	1.1%	3.3%	0%	23.2%	25.0%
Megvii (Face++)	11.9%	9.7%	8.2%	13.9%	32.4%	46.5%
IBM	5.1%	7.4%	8.2%	8.3%	33.3%	46.8%

Table 3. From 2018 Gender Shades Study: Binary-Gender Classification Error Rates on Women by Fitzpatrick Skin Type

For example, when evaluating error rates for the the facial analysis task of binary-gender classification (which does not account for gender nonconforming people, nonbinary people, agender people, and/or transgender people), our 2018 Gender Shades audit showed women with skin types associated with blackness had error rates as high as 47%. In the same study for men with skin-types perceived as white, error rates were no more than .08% in aggregate. The 47% error rate is of note because binary-gender classification has a 50/50 chance of success based on a random guess.

In our follow up 2019 Algorithmic Justice League Actionable Auditing study, my colleagues and I found that even when target companies improved binary-gender classification performance, publicly attributed to improved training data, they still performed better on lighter-skinned than darker-skinned faces, performed better on male-identified faces than female-identified faces, and performed worst on women of color. Even if accuracy disparities are within a few percentage points, differential performance on millions or hundreds of millions of people will impact many individuals. Therefore, to the extent possible, we need to make sure that we cultivate a practice of employing intersectional analysis in how AI is taught, researcherid, and developed.

EROSION OF PUBLIC INTEREST

In researching and remedying with issues around the ethical and societal implications of AI, the public interest must be prioritized ahead of business interests that are incentivized to maximize profitability over other potential outcomes. The risks associated with AI harms and bias threaten trust in government agencies as well as the reputation and product acceptability of influential technology companies who are increasingly funding research and influencing policy discussion around ethics, transparency, accountability, and fairness in AI. Without explicit measures to address conflicts of interests and to protect researchers whose work for the public interest are in tension with private interests, public-private partnerships can

lose legitimacy and critical AI harms research may be discouraged or underfunded. As Harvard Law Professor Yochai Benkler writes in a Nature op-ed on the recent Amazon-NSF partnership:

> "When the NSF lends Amazon the legitimacy of its process for a $7.6-million programme (0.03% of Amazon's 2018 research and development spending), it undermines the role of public research as a counterweight to industry-funded research...Yes, institutions have erected some safeguards. NSF will award research grants through its normal peer-review process, without Amazon's input, but Amazon retains the contractual, technical and organizational means to promote the projects that suit its goals. "[52]

We cannot forget that companies have no obligation to prioritize the public interest and every incentive to use their power and influence to diminish threats to profitability. Earlier this year, I experienced firsthand corporate backlash after publishing a research study alongside, Deborah Raji, an undergraduate researcher at the time. Our research demonstrated that Amazon Rekognition displayed gender and skin-type bias for the task of gender classification. Amazon Web Services' (AWS) general manager of artificial intelligence, Matthew Wood, and vice president of global public policy Michael Punke attempted to discredit the research with verifiably false claims. One false claim was that I had not made my research methodology available. The methodology used for the study stems from my MIT Master's thesis which was made public in 2017. The 2018 peer-reviewed paper that built on that work also made the methods clear and reproducible. The data used for the study is publicly available for non-commercial use. Researchers in companies such as IBM and Microsoft who could not agree to the terms have instead reproduced comparable data and results using the guidelines written in the original paper. The attacks from Amazon prompted over 70 researchers to write an open letter defending the research and calling for Amazon to stop selling their technology to facial recognition.[53]

Dr. Yoshio Bengio, a recent Turing Prize winner and machine learning pioneer, was one of the authors of this open letter who has been particularly vocal about the need to make sure companies do no usurp AI faculty to the detriment of building the academic capacity of the field. We need not only to preserve academic talent, but also preserve space for critical research within AI. Dr. Bengio observed "The fact that company representatives chose to refute the Raji and Buolamwini paper highlights the importance of a rational and open debate, which will hopefully discourage other companies from using similar tactics, and instead encourage them to

---

[52] Yochai Benkler, "Don't let industry write the rules for AI" (May 2019), https://www.nature.com/articles/d41586-019-01413-1

[53] Ali Alkhatib et al., "On Recent Research Auditing Commercial Facial Analysis Technology" (March 2019) https://medium.com/@bu64dcjrytwitb8/on-recent-research-auditing-commercial-facial-analysis-technology-19148bda1832

improve their products appropriately and engage in a constructive dialogue with scientists who work on these issues."[54]

I can attest that as a researcher who has faced corporate hostility for the work I do, seeing the same corporation that publicly attacked my work showing algorithmic bias in one of their controversial products now sponsor government research in AI Fairness is troubling. Without clear mechanisms that address conflicts of interests (perceived or otherwise), I am less inclined to seek NSF funding for the research that falls under AI Fairness. However, the government should be a major source of funding for research that falls in the realm of technology in the public interest which includes funding for AI Fairness research that is given the space to unabashedly speak truth to power.

Though a plethora of problems exist when considering the depth of the ethical and societal implications of AI, we still have time to institute countervailing mechanisms so that the color of your skin or the inferred contents of your character do not limit access to opportunity under the banner of machine neutrality. Recommendations for government and academia to address the concerns explicated above are briefly outlined below and followed with more broadly focused measures:

PRIVILEGED IGNORANCE: as further outlined below increase awareness of AI harms and adopt proven techniques to diversify the field, such as gathering and sharing demographic data, setting public time-bound diversity and inclusion targets; establishing community review boards that provide real-world perspectives and checks

MISLEADING EVALUATIONS: systematically audit benchmarks for demographic and other relevant categories of bias; establish more diverse and inclusive benchmarks; adopthuman analysis of real-world biased outcomes beyond the mere evaluation of models.

PROBLEMATIC DATA FLOWS: *implement* stronger requirements for consensual data use, to minimize the harms of nonconsensual data use by AI researchers and practitioners;require documentation of data collection and classification processes to increase due diligence

SINGLE-AXIS ANALYSIS: where applicable require intersectional analysis in government funded research, establishing intersectional audit norms, require NIST and other government agencies assessing algorithmic performances to conduct intersectional audits and/or establish a partnership with universities or independent certified agencies to conduct such audits

EROSION OF PUBLIC INTEREST: establish fully autonomous funding for ethics, transparency, accountability and fairness research; procurement processes that require all private vendors of AI services to public agencies to comply with ongoing intersectional bias audits; a requirement

[54] Dina Bass, "Amazon Schooled on AI Facial Technology By Turing Award Winner" (April 2019) https://www.bloomberg.com/news/articles/2019-04-03/amazon-schooled-on-ai-facial-technology-by-turing-award-winner

for vendors to submit to community review boards that include members of the most-impacted communities; establish better reporting mechanisms for people to share experiences of harm; decriminalized research in the public interest that is currently penalized by the Computer Fraud and Abuse Act.

Broad Recommendations

INCREASE AWARENESS ABOUT AI HARMS AND BIAS

The public is largely unaware of the ways in which AI shapes their lives, and there are few regulations that require disclosure about the use of the technology. Without awareness about the uses, risks, and limitations of AI, we remain at the mercy of entities that benefit from opaque AI systems, even when they propagate structural inequalities and violate civil rights and liberties. Furthermore practitioners tend to be shielded or removed from the work impacts of AI requiring a shift in how we educate current and future AI developers and researchers. Counteracting the harms of AI and ensuring its benefits are more equitably distributed will require making known existing harms.

Ensure that Computer Science Curriculum from K-12 and post secondary institutions alike addresses issues about the societal and ethical implications of AI and emphasizes that the creators of these systems have an obligation to develop AI in a responsible manner. Examples that are used should be based on real-world occurrences of issues and not theoretical abstractions of hypothetical harms in order to make the issues concrete and stress the need for interdisciplinary knowledge.

Resource public interest technology clinics at degree granting institutions with AI-relevant programs. Such clinics can be modeled on public interest law clinics, so that part of AI education includes a requirement for learning about the real-world consequences of algorithmic harms.

Invest in science communication efforts to make accessible the findings of research studies and results on documenting government testing of AI systems. For example, NIST has been charged with developing AI standards for the United States. The studies produced as an output of these efforts should be presented in a manner where non-domain experts can understand the purpose of the research, the limitations of the methods, and the real-world implications of the results. Researchers receiving government support can be incentivized for making efforts to make their research more accessible. Academic institutions should include course or workshops to help researchers become better communicators of their work.

Promote deeper collaborations between AI researchers and organizations that work most closely with communities that are most harmed by algorithmic inequality. Fund university/community partnerships both to study AI harms on marginalized groups, and also to

do participatory design of AI that is rooted in the needs of marginalized communities. Such collaboration will give a much better contextual understanding of the impact of AI on society, and more importantly enable those who are most impacted by AI harms to be part of the process of counteracting these harms.

Promote creative science initiatives that incorporate the arts and media making to reach broad audiences who otherwise may not encounter research-backed information about existing harms of AI that extend beyond science fiction.

CHANGE RESEARCH & INDUSTRY PRACTICES THAT OBSCURE AI HARMS

Lax research standards plague the field such that critical information about the data used in studies is not collected and/or disclosed,and the homogenous demographic composition of key benchmarks and evaluation norms obscures the potential distribution of harms among different populations. Furthermore, AI products and services are sold with little if any information about potential risks, limitations, and out of context use cases.

Academic institutions and government funding agencies can increase expectations by requiring researchers exploring human-center AI to collect demographic and/or other relevant categorical information as well as document the sourcing, labeling, and interpretation of data collected. Documentation standardization efforts like Datasheets for Datasets [55]and Data Nutrition labels[56] provide starting points for considering what kind of information needs to be collected to inform minimum requirements.

Industry and academics developing AI enabled products or general purpose models should document model performance and provide results to inform stakeholders like organizations considering AI integrations, fellow academics, and the general public. Processes like Model Cards for Model Performance[57] provide a baseline template for considering what kind of information needs to be collected at a minimum so stakeholders can make informed decisions.

INVEST IN DIVERSIFYING THE AI FIELD

The lack of diversity in the field of AI is appalling, particularly considering the wide ranging impact of the output of the technologies that are being developed.

Industry, academia, and government should promote known best practices for addressing diversity gaps. At a minimum transparency in the state of the field is needed which

[55] Gebru et al. “Datasheets for Datasets” (2018), https://arxiv.org/abs/1803.09010
[56] Holland et al. “The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards” (2018), https://arxiv.org/abs/1805.03677
[57] Mitchell et al. “Model Cards for Model Reporting” (2019), https://arxiv.org/abs/1810.03993

will involve gathering demographic data; publishing diversity and inclusion goals; publishing a timeline for reaching those goals; releasing at least annually about progress towards goals.

Support affinity groups that are emerging to address diversity gaps such Black in AI and LatinX in AI.

Provide funding for underrepresented employees, students, and academics to participate in industry events and conferences that may be prohibitively expensive.

CREATE AN AI ACCOUNTABILITY FUND TO SUPPORT CRITICAL RESEARCH AND REDRESS HARMS

Instead of having corporations fund the very research that is meant to keep them accountable, an alternative approach for involving corporations in supporting work in the public interest could be the introduction of an AI Accountability Tax that companies deploying AI systems on a significant portion of the US population must pay. If even Alphabet, Microsoft, Amazon, Facebook, IBM, and Apple tech companies paid .5% of annual revenue, the government raise have $4.4 billion. As a tax, instead of a voluntary contribution through corporate partnership, this money would not be contingent on corporate appetites for engaging with issues of the ethical and societal impacts of AI. Companies like Amazon that have found mechanism to avoid paying corporate taxes, would as a result of their reach and influence still have an obligation to the AI Accountability Fund. Further to support efforts like assess and mitigate AI harms bias, companies architect their tools to enable third-party testing i the public interest that does not incur additional costs for researchers doing this work.

Conclusion

Since AI is being integrated into areas of society including healthcare, education, employment, housing, transportation, and criminal justice that have been shaped by unjust histories and practices, government officials, researchers, and practitioners in the field of AI have an increased responsibility to be especially attuned to the limitations of AI systems that can mask and further systematize structural inequalities regardless of intention. **Algorithmic failures are ultimately human failures that reflect the priorities, values, and limitations of those who hold the power to shape technology.**

We must work to redistribute power in the design, development, deployment, and governance of AI if we hope to realize the potential of this powerful advancement and attend to its perils. We must make sure that the future of AI development, research, and education in the United States is truly of the people, by the people, and for all the people, not just the powerful and privileged.

I look forward to answering your questions,
Joy