RESEARCH FINDINGS

There is a diversity crisis in the AI sector across gender and race. Recent studies found only 18% of authors at leading AI conferences are women,[i] and more than 80% of AI professors are men.[ii] This disparity is extreme in the AI industry:[iii] women comprise only 15% of AI research staff at Facebook and 10% at Google. There is no public data on trans workers or other gender minorities. For black workers, the picture is even worse. For example, only 2.5% of Google's workforce is black, while Facebook and Microsoft are each at 4%. Given decades of concern and investment to redress this imbalance, the current state of the field is alarming.

The AI sector needs a profound shift in how it addresses the current diversity crisis. The AI industry needs to acknowledge the gravity of its diversity problem, and admit that existing methods have failed to contend with the uneven distribution of power, and the means by which AI can reinforce such inequality. Further, many researchers have shown that bias in AI systems reflects historical patterns of discrimination. These are two manifestations of the same problem, and they must be addressed together.

The overwhelming focus on 'women in tech' is too narrow and likely to privilege white women over others. We need to acknowledge how the intersections of race, gender, and other identities and attributes shape people's experiences with AI. The vast majority of AI studies assume gender is binary, and commonly assign people as 'male' or 'female' based on physical appearance and stereotypical assumptions, erasing all other forms of gender identity.

Fixing the 'pipeline' won't fix AI's diversity problems. Despite many decades of 'pipeline studies' that assess the flow of diverse job candidates from school to industry, there has been no substantial progress in diversity in the AI industry. The focus on the pipeline has not addressed deeper issues with workplace cultures, power asymmetries, harassment, exclusionary hiring practices, unfair compensation, and tokenization that are causing people to leave or avoid working in the AI sector altogether.

The use of AI systems for the classification, detection, and prediction of race and gender is in urgent need of re-evaluation. The histories of 'race science' are a grim reminder that race and gender classification based on appearance is scientifically flawed and easily abused. Systems that use physical appearance as a proxy for character or interior states are deeply suspect, including AI tools that claim to detect sexuality from headshots,[iv] predict 'criminality' based on facial features,[v] or assess worker competence via 'micro-expressions.' [vi] Such systems are replicating patterns of racial and gender bias in ways that can deepen and justify historical inequality. The commercial deployment of these tools is cause for deep concern.

[i] Element AI. (2019). Global AI Talent Report 2019. Retrieved from https://jfgagne.ai/talent-2019/.

[ii] AI Index 2018. (2018). Artificial Intelligence Index 2018. Retrieved from http://cdn.aiindex.org/2018/AI%20Index%202018%20Annual%20Report.pdf.

[iii] Simonite, T. (2018). AI is the future - but where are the women? *WIRED*. Retrieved from https://www.wired.com/story/artificial-intelligence-researchers-gender-imbalance/.

[iv] Wang, Y., & Kosinski, M. (2017). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*.

[v] Wu, X. and Zhang, X. (2016). Automated Inference on Criminality using Face Images. Retrieved from https://arxiv.org/pdf/1611.04135v2.pdf.

[vi] Rhue, L. (2018). Racial Influence on Automated Perceptions of Emotions. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3281765.

INTRODUCTION

There is a diversity crisis in the AI industry, and a moment of reckoning is underway. Over the past few months, employees have been protesting across the tech industry where AI products are created. In April 2019, Microsoft employees met with CEO Satya Nadella to discuss issues of harassment, discrimination, unfair compensation, and lack of promotion for women at the company.[1] There are claims that sexual harassment complaints have not been taken seriously enough by HR across the industry.[2] And at Google, there was an historic global walkout in November 2018 of 20,000 employees over a culture of inequity and sexual harassment inside the company, triggered by revelations that Google had paid $90m to a male executive accused of serious misconduct.[3]

This is just one face of the diversity disaster that now reaches across the entire AI sector. The statistics for both gender and racial diversity are alarmingly low. For example, women comprise 15% of AI research staff at Facebook and just 10% at Google.[4] It's not much better in academia, with recent studies showing only 18% of authors at leading AI conferences are women,[5] and more than 80% of AI professors are male.[6] For black workers, the picture is worse. For example, only 2.5% of Google's workforce is black,[7] while Facebook and Microsoft are each at 4%.[8,9] We have no data on trans workers or other gender minorities. Given decades of concern and investment to redress the imbalances, the current state of the field is alarming.

The diversity problem is not just about women. It's about gender, race, and most fundamentally, about power.[10] It affects how AI companies work, what products get built, who they are designed to serve, and who benefits from their development.

This report is the culmination of a year-long pilot study examining the scale of AI's current diversity crisis and possible paths forward. This report draws on a thorough review of existing literature and current research working on issues of gender, race, class, and artificial intelligence. The review was purposefully scoped to encompass a variety of disciplinary and methodological perspectives, incorporating literature from computer science, the social sciences, and humanities. It represents the first stage of a multi-year project examining the intersection of gender, race, and power in AI, and will be followed by further studies and research articles on related issues.

[1] Tiku, N. (2019, Apr. 4). Microsoft Employees Protest Treatment of Women to CEO Nadella. *WIRED*. Retrieved from https://www.wired.com/story/microsoft-employees-protest-treatment-women-ceo-nadella/.

[2] Gershgorn, D. (2019, Apr. 4). Amid employee uproar, Microsoft is investigating sexual harassment claims overlooked by HR. *Quartz*. Retrieved from https://qz.com/1587477/microsoft-investigating-sexual-harassment-claims-overlooked-by-hr/.

[3] Statt, N. (2018, Nov. 2). Over 20,000 Google employees participated in yesterday's mass walkout. *The Verge*. Retrieved from https://www.theverge.com/2018/11/2/18057716/google-walkout-20-thousand-employees-ceo-sundar-pichai-meeting.

[4] Simonite, T. (2018). AI is the future - but where are the women? *WIRED*. Retrieved from https://www.wired.com/story/artificial-intelligence-researchers-gender-imbalance/.

[5] Element AI. (2019). Global AI Talent Report 2019. Retrieved from https://jfgagne.ai/talent-2019/.

[6] AI Index 2018. (2018). Artificial Intelligence Index 2018. Retrieved from http://cdn.aiindex.org/2018/AI%20Index%202018%20Annual%20Report.pdf.

[7] Google. (2018). Google Diversity Annual Report 2018. Retrieved from https://static.googleusercontent.com/media/diversity.google/en//static/pdf/Google_Diversity_annual_report_2018.pdf.

[8] Williams, M. (2018, July 12). Facebook 2018 Diversity Report: Reflecting on Our Journey. Retrieved from https://newsroom.fb.com/news/2018/07/diversity-report/

[9] Microsoft. (2019). Diversity & Inclusion. Retrieved from https://www.microsoft.com/en-us/diversity/default.aspx.

[10] As authors of this report, we feel it's important to acknowledge that, as white women, we don't experience the intersections of oppression in the same way that people of color and gender minorities, among others, do. But the silence of those who experience privilege in this space is the problem: this is in part why progress on diversity issues moves so slowly. It is important that those of us who do work in this space address these issues openly, and act to center the communities most affected.

To date, the diversity problems of the AI industry and the issues of bias in the systems it builds have tended to be considered separately. But we suggest that these are two versions of the same problem: issues of discrimination in the workforce and in system building are deeply intertwined. Moreover, tackling the challenges of bias within technical systems requires addressing workforce diversity, and vice versa. Our research suggests new ways of understanding the relationships between these complex problems, which can open up new pathways to redressing the current imbalances and harms.

From a high-level view, AI systems function as systems of discrimination: they are classification technologies that differentiate, rank, and categorize. But discrimination is not evenly distributed. A steady stream of examples in recent years have demonstrated a persistent problem of gender and race-based discrimination (among other attributes and forms of identity). Image recognition technologies miscategorize black faces,[11] sentencing algorithms discriminate against black defendants,[12] chatbots easily adopt racist and misogynistic language when trained on online discourse,[13] and Uber's facial recognition doesn't work for trans drivers.[14] In most cases, such bias mirrors and replicates existing structures of inequality in society.

In the face of growing evidence, the AI research community, and the industry producing AI products, has begun addressing the problem of bias by building on a body of work on fairness, accountability, and transparency. This work has commonly focused on adjusting AI systems in ways that produce a result deemed "fair" by one of various mathematical definitions.[15] Alongside this, we see growing calls for ethics in AI, corporate ethics boards, and a push for more ethical AI development practices.[16]

But as the focus on AI bias and ethics grows, the scope of inquiry should expand to consider not only how AI tools can be biased technically, but how they are shaped by the environments in which they are built and the people that build them. By integrating these concerns, we can develop a more accurate understanding of how AI can be developed and employed in ways that are fair and just, and how we might be able to ensure both.

Currently, large scale AI systems are developed almost exclusively in a handful of technology companies and a small set of elite university laboratories, spaces that in the West tend to be extremely white, affluent, technically oriented, and male.[17] These are also spaces that have a history of problems of discrimination, exclusion, and sexual harassment. As Melinda Gates describes, "men who demean, degrade or disrespect women have been able to operate with such impunity—not just in Hollywood, but in tech, venture capital, and other spaces where

[11] Alcine, J. (2015). Twitter. Retrieved from https://twitter.com/jackyalcine/status/615329515909156865.

[12] Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016, May 3). Machine Bias. *ProPublica*, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[13] Vincent, J. (2016, Mar 24). Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. *The Verge*. Retrieved from https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist.

[14] Melendez, S. (2018, Aug. 9). Uber driver troubles raise concerns about transgender face recognition. *Fast Company*, Retrieved from https://www.fastcompany.com/90216258/uber-face-recognition-tool-has-locked-out-some-transgender-drivers.

[15] Narayanan, A. (2018). 21 fairness definitions and their politics. ACM Conference on Fairness, Accountability and Transparency. Retrieved from https://www.youtube.com/watch?v=jIXIuYdnyyk.

[16] Vincent, J. (2019, Apr. 3). The Problem with AI Ethics. *The Verge*. Retrieved from https://www.theverge.com/2019/4/3/18293410/ai-artificial-intelligence-ethics-boards-charters-problem-big-tech.

[17] Crawford, K. (2016, June 25). Artificial Intelligence's White Guy Problem. *The New York Times*. Retrieved from https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html.

their influence and investment can make or break a career. The asymmetry of power is ripe for abuse."[18] Or as machine learning researcher Stephen Merity noted at the end of 2017, "Bias is not just in our datasets, it's in our conferences and community."[19]

Both within the spaces where AI is being created, and in the logic of how AI systems are designed, the costs of bias, harassment, and discrimination are borne by the same people: gender minorities, people of color, and other under-represented groups. Similarly, the benefits of such systems, from profit to efficiency, accrue primarily to those already in positions of power, who again tend to be white, educated, and male. This is much more than an issue of one or two bad actors: it points to a systematic relationship between patterns of exclusion within the field of AI and the industry driving its production on the one hand, and the biases that manifest in the logics and application of AI technologies on the other.

Addressing these complexities will take much more than the technically-driven problem solving that has thus far dominated the discussion of gender and race in AI. Our research points to the need for a more careful analysis of the ways in which AI constructs and amplifies systems of classification, which themselves often support and naturalize existing power structures,[20] along with an examination of how these systems are being integrated into our institutions, and how they may be experienced differently on the basis of one's identity. Such research requires looking at gender and race as categories "within which humans think about and organize their social activity, rather than a natural consequence of difference."[21] In short, in studies of discriminatory systems we need to ask: who is harmed? Who benefits? Who gets to decide?

It is critical that we not only seek to understand how AI disadvantages some, but that we also consider how it works to the advantage of others, reinforcing a narrow idea of the 'normal' person.[22] By tracing the way in which race, gender, and other identities are understood, represented, and reflected, both within AI systems, and in the contexts where they are applied, we can begin to see the bigger picture: one that acknowledges power relationships, and centers equity and justice.[23]

[18] Kolhatkar, S. (2017). The Tech Industry's Gender Discrimination Problem. https://www.newyorker.com/magazine/2017/11/20/the-tech-industrys-gender-discrimination-problem.

[19] Merity, S. (2017). Bias is not just in our datasets, it's in our conferences and community. *Smerity.com*. https://smerity.com/articles/2017/bias_not_just_in_datasets.html.

[20] Bowker, G.C. and Star, S.L. (1999). *Sorting Things Out: Classification and its Consequences*. Cambridge: MIT Press.

[21] Harding, S. (1986). *The Science Question in Feminism*. Ithaca: Cornell University Press, p. 17

[22] While race and gender are key axes of identity, and are most commonly considered in discussions of AI bias, it is important to emphasize that they are far from the only identity categories that shape AI systems. For example, as the work of Virginia Eubanks makes clear, class-based discrimination is a particularly thorny challenge, highlighting the ways in which AI systems are entwined with surveillance of the poor. See: Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Punish and Police the Poor*. London: St. Martin's Press. In addition, in partnership with the NYU Center for Disability Studies and Microsoft, AI Now recently hosted a one-day workshop on Disability and Bias in AI. We will be releasing a report summarizing our discussion and examining the ways in which disability studies expand and complicate our notions of AI bias. An examination of disability in the context of AI bias is particularly productive in that it requires us to scrutinize what (and who) constitutes a "normal" body, how aberrance and normalcy are themselves defined (and by whom), how such normative classifications may be mapped onto bodies in different ways at different times throughout an individual's lifetime, and what the consequences of such classifications may be.

[23] For thoughtful treatments of what a justice-oriented data science might look like, and how it differs from data ethics, see: Green, B. (2018). Data Science as Political Action: Grounding Data Science in a Politics of Justice, Retrieved from https://scholar.harvard.edu/files/bgreen/files/data_science_as_political_action.pdf, and Klein, L. and D'Ignazio, C. (2019). *Data Feminism*. Cambridge: MIT Press. Retrieved from https://bookbook.pubpub.org/pub/dgv16l22.