Artificial Intelligence: The Double-Edged Sword

By Massimo Ginella, Cabrillo College
Mentors: Letitia Scott-Curtis and Steve Schessler

As powerful Silicon Valley companies continue to pour millions into the development of new tech, Artificial Intelligence increasingly is at the core of development growth. AI relies on the input of developers and training data in order to “learn”. The predominantly homogenous workforces of these tech companies - 83% men, and more specifically 50% white men - have begun to unknowingly produce Artificial Intelligence models that are often optimized for its developers, therefore negatively impacting marginalized communities. This inadvertent bias produces unvarying training data which causes discrepancies regarding optimization for people of different genders and ethnicities (Harkinson). These newly developed AI models have been known to frequently demonstrate biased results through facial recognition, medical, and recruitment software. Additionally, when algorithms are tested, certain Artificial Intelligence models have demonstrated racial and gender biases which can be reflected in the AI systems entrenched in common technology such as our smartphones, computers, and even vehicles. Silicon Valley tech companies must diversify their work forces further. Recruiting students from diverse colleges and universities and providing intentional scholarship and internship opportunities, will yield positive results in stopping these AI from replicating and spreading racial and sexist stereotypes. Not only this, but simultaneously diversifying the often homogenous training data fed to AI models will yield similar results. Broadening training data collection through data copyright and manual data sorting will prove beneficial in creating a positive change in slowing the development of AI that seem to replicate inappropriate ideals. With these changes, future AI models will yield a higher percentage of positive, and unbiased results due to a fresh set of cultures, experiences, eyes, and minds in the workforce along with diverse training data.

A 2014 article written by Josh Harkinson investigated the diversity of large Silicon Valley tech companies. The results from Harkinson’s study demonstrated that “83 percent of the tech jobs are held by men, and 94 percent of those workers are white or Asian” (Harkinson). Along with this, companies such as Google, Facebook, and Microsoft hold similar statistics in which more than 80% of the worker population is made up of two main categories (Ariella) (Figure 1). The perspectives of people of different cultures, genders, and ethnicities can serve as an effective method of fabricating products that are optimized for everyone, therefore lowering the risk of discriminatory products being produced.

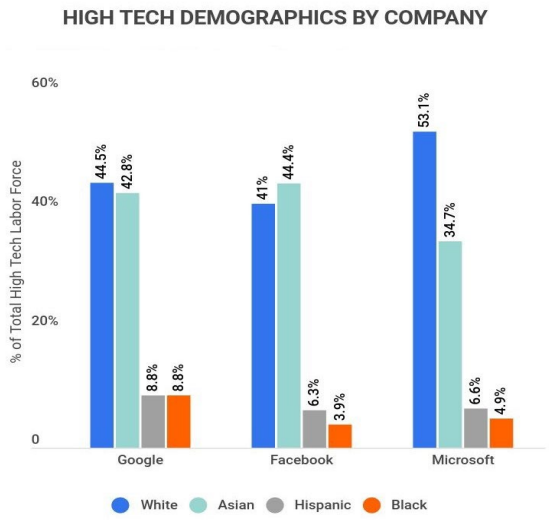

Figure 1: Diversity statistics of the largest Silicon Valley Companies (Ariella

Often when a diverse workforce is neglected, the product created by that workforce may end up representing only a single homogenous group of people. A good reference of this phenomena would be an occurrence within the company known as *BAND-AID*, parented by Johnson & Johnson. “While the product line has expanded to incorporate new shapes and sizes, the “flesh colored” Band-Aid remained uniquely pink in hue until 2020, when civil unrest surrounding anti-Black police brutality forced corporations like J & J to take action” (Press). This lack of diversity from large Silicon Valley tech companies is one of the leading factors in the production of AI that closely expresses insensitive behavior. As of 2020, out of Silicon Valley's population

of 2 million residents, 140,000 people claimed they have worked IT jobs (He). When such a large workforce is made up of a narrow selection of individuals, the production of AI is often optimized solely for that specific group of people as it is only tested in scenarios where these individuals with specific characteristics would utilize it. This is exactly the root of the issue with the AI produced over the last two decades: it remains optimized primarily for those who develop it and does not account for a variety of skin tones, features, and genders.

Before journeying further into how AI is capable of replicating human bias due to invariable workforces, it's important to understand the process in which one of these models is developed. There are five broad stages to development, data collection, data cleansing, training, testing, and finally, release. The data collection period is heavily governed by what type of data is needed to allow the AI model to function. For example, if it's a language model like ChatGPT, its primary training data is text documents so it can recognize patterns in our dialect and make accurate predictions for its generative nature; if it's an image recognition model such as FaceNet, it will likely be images of humans so that it can find commonalities in the human facial structure and therefore recognize faces accurately. Once a type of data is chosen for training and that data type is gathered in mass quantity, development teams will cleanse that data for any possible discrepancies, inaccuracies, or incorrect information. In this case, high quality data means high quality performance. Once the data has been gathered and cleansed, the development teams will feed this data to their AI model as training data, therefore allowing it to make predictions of the data and optimizing its performance so it can make more accurate conclusions. When the AI model has been fed enough data and is capable of accurate decision making, the model will be put through (sometimes) rigorous testing phases in which the AI is optimized through calculus topics to make the closest possible predictions for its final decision-making process. Finally, if all is well, the model will be released for whatever real world use it may have. The main area of focus within this five-stage process that leads to AI bias is both the data gathering, and data cleansing phases. Without training data, an AI model has nothing to work with, and therefore no conception of what task it must be optimized to accomplish. The quality and diversity of data of which an AI model is given heavily affects its decision making.

Recently, a common AI used in medical centers to survey patients for possible illnesses and refer them to programs that can give them the care they need has frequently demonstrated results that mirror and compound the issues of medical racism perpetuated in the past by medical institutions and insurance companies. This AI, which is utilized and has helped manage care for approximately 200 million people in the US annually, has been prioritizing the health of lighter skinned patients over those with darker skin tones (Ledford).

> *When Obermeyer and his colleagues ran routine statistical checks on data they received from a large hospital, they were surprised to find that people who self-identified as black were generally assigned lower risk scores than equally sick white people. As a result, the black people were less likely to be referred to the programmes that provide more-personalized care. (Ledford)*

Researchers later discovered that the cause behind the AI demonstrating racist ideals was due to its strange decision-making skills through the use of risk scores (Ledford). In medical fields, risk scores are a metric utilized to estimate a patient's likelihood of future illness, injury, or medical events given their current health status. With this being said, medical Artificial Intelligence models have been known to assign higher risk scores to patients with lighter skin tones than those with darker skin tones, therefore preventing those who may have darker skin pigmentations from receiving the proper medical care they need, while those with lighter skin pigmentations are labeled as higher priority. With this, an inadvertent hierarchy is created in which discrimination is a key factor in the diagnosis of a patient, therefore possibly exposing millions of people to risk of future medical complications due to misdiagnosis. In this case, there is not a clear cause for this behavior, but it could be a mirroring effect of racism within medical fields, as it is often a strong misconception that people of darker skin tones could carry lesser risk towards certain health conditions than those of lighter skin tones. This could be a direct cause of faulty training data being fed to this particular Machine Intelligence

model in which the data could be composed of misleading medical documentation with discriminatory ideals, therefore leading this model to arrive at prejudiced conclusions from its input data. With this being said, it's entirely possible that the cause of this AI model's behavior is due to inaccurate and poorly cleansed training data, therefore causing it to reach prejudiced conclusions, and can lead to health risks of millions. However, the training data that is fed to AI models in development is often kept behind closed doors, therefore resulting in only speculation towards what could be causing bias in this specific AI model.

However, this is not the only case in which AI has shown questionable results over the color of one's skin. Certain facial recognition softwares have been commonly known to be more successful in detecting the faces of white males over any other gender or skin color, and when presented with anything other than a white male, the results are unsurprisingly inaccurate. Founder of the Algorithmic Justice League, MIT graduate, and writer Joy Buolamwini relates, “I experienced this firsthand, when I was a graduate student at MIT in 2015 and discovered that some facial analysis software couldn’t detect my dark-skinned face until I put on a white mask. These systems are often trained on images of predominantly light-skinned men” (Buolamwini). While Buolamwini was testing facial recognition AI developed by Amazon, Microsoft, and IBM on images of black women such as Oprah Winfrey, Michelle Obama, and Serena Williams, she found that every single facial recognition AI was incorrect in its attempt to identify the race and gender of the individual in the photograph it was provided. When the AI system was run, the photographs of these women were marked as “appears to be male 76.5%” and “a young man wearing a black shirt, confidence: 0.9350064” (Buolamwini). The root of this AI marking the faces of popular black women as male is due to the lack of diversity behind the AI development teams at companies such as Amazon, Facebook, Microsoft, and so on. When images of light-skinned males is the primary source of training data, the results the AI model will output is destined to be biased when tested on non-light-skinned individuals. This is often due to biased training data being fed to AI models during the development phase. Due to the frequency of homogenous workforces in Silicon Valley. For example, if a software development team is made up of mainly people with primarily lighter skin pigmentation, the training data (in this case it would be images of faces for training a facial recognition model) that is picked could be mainly comprised of people who bear a similar resemblance to those in the workforce, therefore only allowing the AI model to receive proper training with people who look similar to those who created it, and perform poorly towards those who don't bear a similar resemblance. Word of these AI has been getting more attention recently, and has led computer scientists to investigate the common algorithms that are utilized in these popular AI systems. One of these models that is being researched is named Contrastive Language-Image Pretraining, also known as “CLIP.” CLIP is an artificial intelligence deep learning framework that was created by OpenAI. Its use as a tool is for connecting an input of images with text in order to match images with language descriptions (Verma). Researchers fed 62 different input commands of mainly text and images into CLIP as a way to test possibly biased results. “When researchers asked robots to identify blocks as “homemakers,” Black and Latina women were more commonly selected than White men, the study showed. When identifying “criminals,” Black men were chosen 9 percent more often than White men” (Verma). The researchers investigating CLIP would later do several more tests for images of people to fill the roles of doctors and janitors, both of which returned biased results as women were less likely to be doctors, and Latino men were more likely to be janitors (Verma). Due to the fact that the algorithms written for AI are often difficult to understand, and even more difficult to gain access to, this leads to an issue in which AI models like CLIP cannot be fixed so easily. If more AI language models such as CLIP were to be utilized in more common technology, additional issues could possibly arise. Verma includes one such scenario:

> *Imagine, he said, a scenario when robots are asked to pull products off the shelves. In many cases, books, children’s toys and food packaging have images of people on them. If robots trained on certain AI were used to pick things, they could skew toward products that feature men or White people more than others, he said. (Verma)*

All of these AI models share one thing in common: training data. Without reference data, Machine Intelligence is useless. In these cases, the training data gathered by large tech companies with low diversity rates has a direct correlation to biased output from their Artificial Intelligence models. This could negatively impact our society in the future through promoting racist and sexist stereotypes through common mediums. Many of these algorithms boil down to the testing phase, in which mainly light skinned male individuals are in charge of what data the AI model is trained on. If the workforces of Silicon Valley tech companies were to be more diverse, it would be much easier to spot discrepancies in training in varying data so that a wider variety of said data can be utilized for more widely optimized results.

As most large tech companies receive a multitude of applications every year, it is quite laborious and cost intensive to support a large board of recruiters. In this case, it has become more frequent for these organizations to utilize recruitment AI to assist with the hiring process. These models are capable of scanning the database of an applicant for information revolving around their social media, job boards, and resumes. Based on the information that is associated with an applicant, they are assigned risk scores regarding how qualified they are for the job they apply for. While this process can help automate the recruiting process and therefore slim down the applicant fields for those who will move to more advanced application stages, these recruitment AI models have shown negative results in often removing individuals from the applicant pool due to factors out of their control.

> *But automatic tools in this area have exhibited worrying biased behaviors in the past. For example, Amazon's recruiting tool was preferring male candidates over female candidates [11]. The access to better job opportunities is crucial to overcome differences of minority groups. However, in cases such as automatic recruitment, both the models and their training data are usually private for corporate or legal reasons. (Peña, Serna, Morales, Fierrez, 1) (Figure 2).*

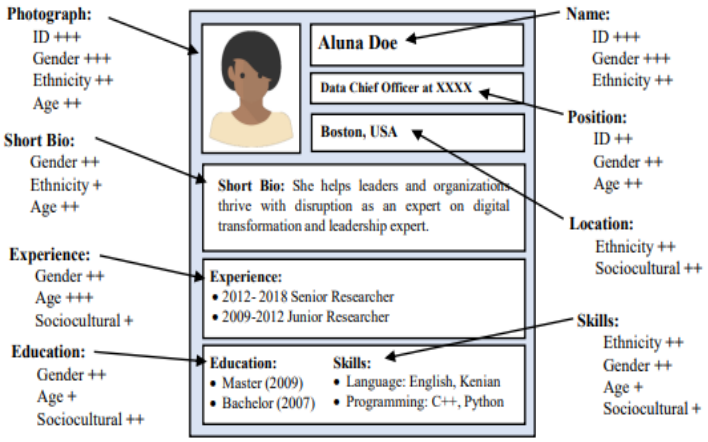

Figure 2: Information on a resume showing sensitive data.(+++ = high, ++ = medium, + = low) (Peña, Serna, Morales, Fierrez, 2)

Recruitment AI have been known to nitpick sensitive data such as someone's gender, place of origin, ethnicity, and age and utilize this information against them as a way to determine whether or not they are qualified for the job they are applying for. This could be occurring due to the fact that the AI model compares the information about an applicant and associates it with the information of those who hold the job they apply for. For example, since the tech industry has been known to be dominated by men, a recruitment AI may have been trained on some data suggesting this, therefore removing female applicants due to the fact that it's "out of the ordinary" for a woman to hold a job position in tech. This behavior only makes it more difficult for unique applicants to be considered for a job they are applying for, and worsens the likelihood of eliminating the homogenous culture of tech workforces.

Given many instances of AI replicating human bias, the central question at hand is how these occurrences are even possible. Many people perceive Machine Intelligence to be inherently neutral, when this is obviously not the case. With this in consideration, we must ask how a non-neutral stance from a machine could manifest. In short, the behavior of these machines is a direct result of human error. The lack of diversity compounds these errors and leads to both poorly selected, and poorly cleansed homogenous training data which is used to train an Artificial Intelligence model. This in turn leads to a negative effect on its neural networking algorithms, better understood as its decision-making abilities. This idea becomes evident in the form of poor optimization in facial recognition models, primarily resulting from a workforce and training data dominated by individuals who are

predominantly light-skinned males. With all three of these errors in play, the output of these AI systems is destined to exhibit some sort of bias. As this is the situation, we currently find ourselves entrenched in, the questions from this point on are how we can mitigate this issue and have there been any initiatives taken to do so?

Although the invariability of tech workforces has been an issue for quite some time, this is not to say that nothing has been done to push for change. During the height of the Black Lives Matter (BLM) movement of 2020, protesters began to put pressure on large tech corporations due to their seemingly unwilling nature to diversify their workforces. Because of the pressure ensued upon companies like Facebook and Google, these tech giants both donated hefty sums of money in support of the BLM movement and pledged to increase underrepresented leadership by 30% (Sor). However, hardly anything has changed regarding the promises to diversity workforces, therefore developing suspicion that these companies may have donated and promised diversity solely for the purpose of good public relations.

> *Blendoor reports that Black and Hispanic workers made up just 4.7% and 6.8% of the industry in 2021, according to its analysis of most recent information from 240 companies. It's less than a 1% increase from what was recorded in 2014 — far from what the equitable world tech leaders envisioned six years ago. (Sor)*

When it comes to large companies, it is often common to hear more about what they plan to do, rather than what they have actually done. The reason that these companies use this tactic is to gain favor from the masses and hope they eventually forget about if there was any initiative taken by the company to solve a certain issue. Although there are certain corporations that seem more hesitant to diversify, companies like Twitter and Intel have been quite successful in diversifying their workforces. "From 2017 to June of this year, the company has nearly doubled the number of Black and Latino employees in its overall workforce, and more than doubled the number of Black employees in leadership" (Sor). Twitter's ability to diversify so rapidly is proof that tech companies of similar size can follow in their footsteps. If most companies aim to mimic Twitter and copy the strategies they have implemented to diversify their workforce, Silicon Valley could see an explosive growth in representation. However, this is not yet the reality and there is still a long way to go before this can be completed.

A way to push these corporations to diversify would be through Executive Order 14035, which was signed by President Joe Biden after taking office in 2021 (Baldwin). Executive Order 14035 pushes for "the practice of including the many communities, identities, races, ethnicities, backgrounds, abilities, cultures, and beliefs of the American people, including underserved communities" into the American workforce (whitehouse.gov). Executive Order 14035 could prove to be a useful tool in forcing Silicon Valley companies to diversify under Federal Law. If tech companies continue to show little to no inclusion in the coming years, with enough support, more attention could be brought to this issue. This order could possibly signify a change of the workforce of these companies under Federal Law. The use of biased recruitment AI models is a direct violation of this order as discrimination of applicants due to sensitive or unrelated data has been a common theme in narrowing down applicant pools.

If these companies were to consider diversifying in the near future, there are numerous ways in which they could approach the hiring process to recruit strong, intelligent, and diverse candidates. One method would be to recruit from diverse educational institutions as it is common for university graduates to seek jobs according to their major straight out of college. According to an article written by Josh Moody containing lists and statistics of the most diverse universities in the US, some colleges and universities which could possibly be strong candidates in this department would be University of Houston, Georgia State University, and Texas Woman's University, all with diversity indexes of 0.73 (Moody). These three universities demonstrate high diversity indexes and hold GPA averages that are higher than the average college student. On top of this, Texas Woman's University has an overall female population of 87%. Silicon Valley companies would have no

issues in finding intelligent, qualified people of all different walks of life if they begin their recruitment process with universities with high diversity indexes such as the ones listed.

Another great way to grow the staff of these companies would be to create intentional scholarship and internship opportunities to under-represented communities. Given the enormous annual revenues of these companies averaging at around one hundred billion dollars per year, there must be plenty of grant and scholarship money to be passed around to communities that are in need of academic tools to train their students such as laptops, textbooks, and general necessities (companiesmarketcap.com). With this in mind, more students being enrolled in schools with this money means a possible increase in these students' interests in STEM related fields, therefore broadening and diversifying the range of people in the STEM world and boosting the amount of future applications for the companies that supported them. As those that are being supported would be from underrepresented communities, hiring these individuals would bring entirely new perspectives and experience to the workforces, therefore diversifying the workforce, likely mitigating homogenous and poorly sorted training data, and lessening the chances of biased output from AI being developed.

While diversifying the workforces of massive tech companies is essential to the mitigation of AI bias, it is also crucial that quality of the data that is fed to AI models during their training phases be of high priority as well. Invariable training data that is used to train AI will inevitably lead to bias, so it is of utmost important to gather as much unique and trustworthy data as possible to minimize the potential of discriminatory outputs. One way that this can be done is through data copyright. The core concept of data copyright is that individuals and entities, such as data analysts, scientists, market researchers, and survey researchers who collect and curate data sets can ensure fair use of the data they gather. This means that anybody who wishes to obtain their collected data must do so through purchase of a licensing agreement with the holder. As tech companies with large budgets seek to purchase data from individual holders, a large incentive is created for individual holders to gather as much reliable, diverse data as possible. With such a large influx of data holders willing to provide diverse data sets to tech companies under fair use, this can lead to a plethora of training data available for AI development, therefore allowing for more favorable outcomes (Levendowski). With this large influx of data, there may be some discrepancies within the data that needs surveying. As AI is used at times to sort through training data sets, while it may be more time consuming and cost intensive, it may be a better idea to have developers manually sort through the data that is provided to avoid possible AI bias in the data cleansing process. This of course requires a non-homogenous workforce in order for effective bias recognition, but it may yield positive results for high quality data being used for training. These two ideas could potentially mitigate the amount of biased data that is used to train future Artificial Intelligence models.

Silicon Valley has a diversity problem, and due to this lack of representation, AI trained on suboptimal, homogenous training data are being created without knowledge of the development teams producing them. These AI have caused emotional, financial, and physical damage in medical fields, facial recognition, and within the job application process. These major Silicon Valley companies must be forced to increase employment rates through recruiting students from diverse colleges and universities, grants and scholarships, and mitigating the use of recruitment AI. Not only this, but the quality of data that is being used to train AI should be assessed by utilizing tools such as data copyright, and manual data sorting for top of the line, diverse data. If success is to be reached in diversifying the workforces of tech giants, the chances of biased training data will reduce, and future AI models will yield a higher percentage of positive, and unbiased results due to a fresh set of cultures, experience, eyes, and minds in the workforce.

Bibliography

Ariella, Sky. “25+ Telling Diversity in High Tech Statistics [2022]: Tech Demographics + Trends.” *Zippia 25+ Telling Diversity In High Tech Statistics 2022 Tech Demographics + Trends Comments*, Zippia, 8 Nov. 2022, https://www.zippia.com/advice/diversity-in-high-tech-statistics/.

Baldwin, Kelsi. “Federal Laws Protecting Diversity in the Workplace.” Mc, 12 Aug. 2022, https://mclawreview.org/2022/08/12/federal-laws-protecting-diversity-in-the-workplace/.

Buolamwini, Joy. “Artificial Intelligence Has a Racial and Gender Bias Problem.” *Time*, Time, 7 Feb. 2019, https://time.com/5520558/artificial-intelligence-racial-gender-bias/

Doty, Chris. “A Beginner's Guide to Machine Learning.” RapidMiner, 6 May 2022, https://rapidminer.com/blog/beginners-guide-to-machine-learning/.

Ganel, Tzvi, et al. “Biases in Human Perception of Facial Age Are Present and More Exaggerated in Current AI Technology.” *Scientific Reports*, vol. 12, no. 1, 2022, pp. 22519–22519, https://doi.org/10.1038/s41598-022-27009-w.

Harkinson, Josh. “Silicon Valley Firms Are Even Whiter and More Male than You Thought.” *Mother Jones*, 29 May 2014, https://www.motherjones.com/media/2014/05/google-diversity-labor-gender-race-gap-workers-silicon-valley/#:~:text=For%20example%2C%2083%20percent%20of,workers%20are%20white%20or%20Asian.

He, Eric. “Silicon Valley Still Dominates U.S. Tech Industry, Study Finds.” *Palo Alto, CA Patch*, Patch, 23 Aug. 2021, https://patch.com/california/paloalto/silicon-valley-still-dominates-u-s-tech-industry-study-finds.

Hunkenschroer, Anna Lena, and Christoph Luetge. “Ethics of Ai-Enabled Recruiting and Selection: A Review and Research Agenda - Journal of Business Ethics.” *SpringerLink*, Springer Netherlands, 8 Feb. 2022, https://link.springer.com/article/10.1007/s10551-022-05049-6.

Levendowski, Amanda. “How Copyright Law Can Fix Artificial Intelligence’s Implicit Bias Problem.” *Washington Law Review*, vol. 93, no. 2, 2018, pp. 579–630.

Ledford, Heidi. “Millions of Black People Affected by Racial Bias in Health-Care Algorithms.” *Nature News*, Nature Publishing Group, 24 Oct. 2019, https://www.nature.com/articles/d41586-019-03228-6.

Moody, Josh. “See the Most Diverse National Universities - US News & World Report.” See the Most Diverse National Universities, *US.News*, 5 May 2021, https://www.usnews.com/education/best-colleges/slideshows/see-the-most-diverse-national-universities.

Peña, Alejandro, et al. Bias in Multimodal AI: Testbed for Fair Automatic Recruitment. IEEE CVPR Workshop on Fair, Data Efficient and Trusted Computer Vision, 2020, https://doi.org/10.48550/arxiv.2004.07173. https://arxiv.org/pdf/2004.07173.pdf

Press, Sara. “A Band-Aid on Systemic Racism.” *S Y N A P S I S*, 9 Dec. 2021, https://medicalhealthhumanities.com/2021/12/09/a-band-aid-on-systemic-racism/.

Sor, Jennifer. “Silicon Valley Pledged to Become More Diverse. A Year Later, Has Anything Changed?” *San Francisco Chronicle*, 30 Aug. 2021, https://www.sfchronicle.com/tech/article/Silicon-Valley-pledged-to-become-more-diverse-A-16414178.php.Silicon Valley pledged to become more diverse. A year later, has anything changed?

“Top Publicly Traded Tech Companies by Revenue.” *CompaniesMarketCap.com - Companies Ranked by Market Capitalization*, https://companiesmarketcap.com/tech/largest-tech-companies-by-revenue/.

Verma, Pranshu. “These Robots Were Trained on AI. They Became Racist and Sexist.” *The Washington Post*, WP Company, 19 July 2022, https://www.washingtonpost.com/technology/2022/07/16/racist-robots-ai/.