Whistleblowing Stochastic Parrots: A Sense of Artificial Intelligence Ethics and Advocacy

Alice Qiang
Tulane University, New Orleans, Louisiana, USA

Abstract: On December 2nd, 2020, former Google Artificial Intelligence Researcher Timnit Gebru announced that she was fired from Google after she refused to retract a paper that called out the social, environmental, and bias risks of large language models. She raised awareness about biases in artificial intelligence (AI), which sparked conversations about the use of AI within policing systems, given the possibility for disproportionate harm marginalized communities could face from these systems. Her paper's key contention was that AI centered on large language models, systems—like OpenAI's ChatGPT and Google's newly created PaLM2—that extracted harmful amounts of data, endangering the environment and pulling from biased sources. Armed with the knowledge of whistleblowing risks, Timnit was faced with an important decision: should she go public about the circumstances of her firing? What if she pushed back and exposed Google's research culture and their mistreatment? What risks would she face in doing so? How would the public react?

Introduction

With fiery eyes, Dr. Timnit Gebru opened Twitter and typed out a sentence that, being in the prime of her career, she never thought she would be writing: "I was fired by @JeffDean for my email to Brain Women and Allies. My corp account has been cut off. So, I've been immediately fired." (Gebru 2020, n.p).

She was fired after publishing a research paper, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" that discussed potential racist protocols within Google's new PaLM2 Artificial Intelligence (AI) algorithm. Gebru, an established AI Researcher, was not unfamiliar with speaking out against the harm of reckless data usage for large language models. As a previous co-lead of Google's Ethics in AI division subcommittee, Timnit strived for ethical practices among the data systems of large Silicon Valley tech giants. She serves as one of the most high-profile Black women within the new field of Ethical AI and Technology that seeks to advocate for fairness, elimination of bias, and responsible use of large language models (LLMs).

Born and raised in Addis Ababa, Ethiopia, Timnit Gebru fled an outbreak of political violence in her homeland as a teenager. She was 15 when Ethiopia first went to war with Eritrea, forcing her to find refuge first in Ireland and then in the United States (Rocha 2023). When she first arrived in the outskirts of Boston, Massachusetts, she and her family were immediately met with racist actions. An agency boss told her mother, an economist in their hometown, to get a job as a security guard because, "who knows whatever degree you get, you are from Africa" (Allison 2022). Her teachers constantly belittled her academic prowess, refusing to place Timnit in advanced classes because "people like [her]" always fail (Allison 2022). Years later, a pivotal experience

with the police put her on the path to analyzing race within technology. She recalls calling the police after she and her friend, another Black woman, were assaulted in a bar in San Francisco. When they arrived, her friend (the victim) was arrested and put in jail, an example of "blatant racism" (Perrigo 2022). In an interview, she reflects "That was the scariest encounter I've ever had in the US" (Perrigo 2022). Gebru had been attacked by a group of guys, and nobody helped them. Gebru recounts being afraid and that her experience was the scariest thing to see, being strangled and yet having people simply walk by and just look at them.

The police had "accused Gebru a number of times and kept telling Gebru to calm down" (Harris 2023). Rather than accurately assessing the situation, the police escalated it. Her experience reaffirmed the systemic discrimination present within law enforcement, leading Gebru to analyze all the systems she was a part of later in her life. This included technology, politics, and the research field of artificial intelligence.

Educational History and Academic Background

In 2004, Dr. Timnit Gebru began her career in engineering, working in software development as an intern and then as a hardware engineer full-time in 2005. She worked at Apple until 2013. At Apple, she designed circuits and signal processing algorithms for various products, including the iPad. She then pursued graduate studies at Stanford University from 2008 to 2017, earning her M.S., and Ph.D. in Electrical Engineering at Stanford's Artificial Intelligence Laboratory. During her Ph.D. program, she conducted research in device physics, optics, and signal processing, and later focused on fine-grained object recognition. Fine-grained object recognition is a computer vision task that involves identifying and distinguishing between objects that are visually very similar, such as different breeds or models of cars. (Krause, Gebru, Fei 2013). Her work combined technical approaches with a sociological perspective, studying correlations between publicly available image sets and factors such as voting patterns and socioeconomic status (LinkedIn; Thomas 2020).

At Stanford, Gebru analyzed millions of street images using AI algorithms (such as Google Maps' Images) to collect information about cars, which she used to make predictions about neighborhood income, voting patterns, race, and education levels. This research, published in the *Proceedings of the National Academy of Sciences*, brought her early recognition for studying biases across disciplines (Thomas 2020). After earning her Ph.D., she completed a postdoctoral fellowship at Microsoft Research's Fairness, Accountability, Transparency, and Ethics (FATE) lab. There, she collaborated with Dr. Joy Buolamwini, founder of the Algorithmic Justice League (AJL) and computer scientist, to conduct a widely cited study auditing commercial facial recognition software, which found significantly higher error rates in identifying dark-skinned women compared to other groups (Rocha 2023). At the FATE lab, Dr. Joy continued her research from the MIT Media Lab, where she worked to identify bias in algorithms and develop the practices for accountability during their design. Together, Dr. Joy and Gebru wrote the paper, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification" (Buolamwi and Gebru 2018), was written in the hopes to improve public transparency about AI models' biased tendencies in its facial recognition and bring more understanding to a convoluted field (Team 2018).

While at the FATE lab, Gebru also attended a 2018 Fairness and Accountability conference where she was interviewed by the *MIT Technology Review* about approaches to mitigating bias in AI systems. She emphasized the importance of diversifying datasets and including a range of annotations related to race, gender, and age as ways to counteract algorithmic bias (Snow 2018).

Her early work auditing biased facial recognition systems went on to influence real-world regulations and helped shape emerging standards and practices to address ethical issues in AI datasets and models.

AI and Anti-Racist Advocacy

After working at the FATE lab, Gebru knew she wanted to continue to apply her skills and expertise to bring ethics into the field of artificial intelligence. Gebru was inspired by ProPublica's 2016 investigation into predictive policing, which detailed how courtrooms across the U.S. were adopting software that claimed to predict the likelihood of defendants reoffending in the future (Perrigo 2022). The software was then used to advise judges during sentencing. By looking at actual recidivism rates and comparing them with the software's predictions, ProPublica found that not only was the software wrong, but it was also dangerously biased, more likely to rate Black defendants who did not reoffend as "high risk" and rate white defendants who did reoffend as "low risk"(Perrigo 2022). The results showed that when an AI system is trained on historical data reflective of systemic inequalities, the system will replicate those inequalities in its predictions. When Gebru read this story, she reflected on her own experiences with racism in the police system and the overwhelming lack of diversity in the AI world. She was very concerned about the future of AI – not because of the risk of rogue machines taking over, but because of the homogenous, one-dimensional group of men who were currently advancing the technology (Perrigo 2022). The lack of diversity among those creating and overseeing AI systems limits the range of perspectives shaping the technology. As a result, algorithms risk being designed in ways that exclude or overlook the experiences, ideas, and needs of certain groups. Because of this, Gebru worried about the consequences that groupthink, insularity, and arrogance would be created in the AI community (Harris 2023).

At a major AI conference in Montreal, she experienced the exclusionary homogeneity of the field. While attending a Google party, she was openly harassed by a group of white men—one of whom kissed her without consent while another took a photo (Harris 2023). Gebru felt frozen and shocked, and when recalling the event a few years later in an interview, emphasized that she had done nothing wrong to invite such behavior. At the time, conference organizers had minimal safeguards in place and later acknowledged that their code of conduct has "since been elaborated" and that they added "a new one-stop contact point for concerns and complaints" (Harris 2023). When she returned to the same conference the following year, Gebru began counting the number of Black attendees. Out of 8,500 delegates, she found only six people of color. The limited diversity she observed at the conference only further confirmed her fears of racial discrimination.

In 2018, Gebru joined Google with the hope of carving out a space where marginalized voices could be protected and supported. Upon the time of her hiring, she was the first Black woman to be a research scientist. And even after her leaving Google, there were only 2 other Black women out of hundreds of research scientists (Hao 2020). Yet from the beginning, she encountered structural barriers and a lack of meaningful protection for AI workers. These protections would look like creating a space where workers could speak up honestly—without fearing for their jobs, their pay, or the way their coworkers and supervisors might treat them afterward. She reflected that large companies like Google seemed resistant to cultural change, explaining: "Okay, maybe I can carve out a small piece… that is safe for marginalized groups … [but] the moment you push a little hard, you're out." As she put it, those who survive often do so only by staying silent (Brown 2021).

During this time, tech companies in Silicon Valley were pouring colossal amounts of

money into a previously obscure field of AI, called machine learning (Perrigo 2022). The idea was that with enough data and processing power, they could teach computers to perform a wide array of tasks, like speech recognition, identifying a face in a photo, or targeting people with ads based on their past behavior. For decades, most AI research had relied on hard-coded rules written by humans and could never address such complex tasks on a wide scale. In 2018, technological companies were feeding computers enormous amounts of data through internet and smartphone evolutions—and use high-powered machines to spot patterns within those data (Perrigo 2022). These new frontiers in artificial intelligence could not only unlock substantial human progress but also generate billions of dollars in profit (Perrigo 2022). In fact, machine learning today remains the basis for many lucrative businesses of the 21st century, powering Amazon's recommendation engines and warehouse logistics and Google's search and assistant functions. AI has surpassed beyond mere assistantship and the answering of simple questions, offering promises to transform the future with AI lawyers that can provide affordable legal advice or AI doctors that can diagnose patients in seconds (Perrigo 2022).

The increased reliance on artificial intelligence worried Gebru. In her time at Google, Gebru continued to publish research papers that described the harms of a reliance on artificial intelligence. The topics ranged from using Google Maps to estimate the demographics of neighborhoods to analyzing a comparison of algorithmic fairness approaches under real-world constraints (Walsh 2022). But, it wasn't until Gebru was told by Google to retract her paper from a conference that suddenly propelled her into the spotlight.

The Paper

Dr. Gebru's paper, "On the Dangers of Stochastic Parrots: Can Language Models be Too Big?" laid out the risks of large language models, given that AI data systems train on staggering amounts of data and have grown increasingly large popular within the last three years (Hao 2022). Despite this, they carried four main risks of large language models: their environmental and financial toll, their tendency to absorb harmful biases, their inability to understand cultural and linguistic nuances, and the growing power imbalance favoring wealthy tech companies. She explained that training these massive systems requires staggering amounts of data and energy, creating significant carbon emissions. Citing a 2019 study by Emma Strubell and her collaborators, Gebru noted that training just one model could produce over 626,000 pounds of carbon dioxide—the same as five cars over their entire lifetimes (Hao, 2020). To her, this illustrated how the costs of AI development are borne unequally: powerful corporations reap the benefits while marginalized communities face the environmental consequences. Beyond the ecological impact, Gebru also worried about what happens when these systems are fed the entirety of the internet—racist, sexist, and abusive language included. Without understanding the cultural shifts that shape new movements like #MeToo or #BlackLivesMatter, these models can't truly grasp how language evolves to challenge injustice. As her coauthor Emily Bender explained, the goal was to "take stock of the landscape of current research in natural-language processing," (Hao 2022) and to point out that even the people building these systems may not fully understand the scale or complexity of the data they're using. While Gebru acknowledged that some researchers were working to reduce bias and environmental costs, she ultimately believed that these efforts hadn't gone far enough.

At Google, researchers are required to submit their papers in an internal review before submitting them to external conferences (Hao 2022). This standard practice in many tech companies ensures proprietary or sensitive information is inadvertently disclosed to the company's

management before publishment. The paper was submitted and had been approved by some parts of the internal review chain. After the paper was submitted to an academic conference, it was escalated within Google to higher levels of management (O'Leary 2020). Senior Google executives including Megan Kacholia, the vice president of Google Research, raised objections after discovering the paper from internal review, demanding that Dr. Timnit Gebru and her colleagues either retract the paper from the conference or retract her names and her Google co-authors from the paper.

Thrown in a sudden situation: resignation or termination?

Timnit Gebru's frustration soon turned to disbelief when the situation took an even stranger turn. Her manager, Samy Bengio, informed her that Google executive Megha Kacholia had sent him a document outlining supposed flaws in her paper—but instructed him not to share it with her directly, only to read it aloud. The document, Gebru later recalled, offered no concrete feedback, only vague claims that the paper treated its topics "too casually" and cast new AI technologies in a negative light. Confused and disturbed by the lack of transparency, she spent her Thanksgiving drafting a six-page response titled *"Addressing Feedback from the Ether at Google."* In it, she defended her work and asked for constructive guidance on how to revise the paper rather than suppress it (Simonite 2021).

That weekend, Gebru left for a planned cross-country road trip. While driving through New Mexico, she received an email from Kacholia demanding confirmation that the paper would either be withdrawn from conference consideration or stripped of Google's affiliation. Gebru, feeling silenced, tweeted about "censorship and intimidation" against AI ethics researchers. The following day, she sent two emails that would ultimately end her career at Google (Simonite 2021).

The first, sent to Kacholia, offered a compromise: she would remove her name from the paper if Google explained who had reviewed it and committed to a more transparent research process. If not, she planned to resign after ensuring her team was stable. The second email—sent to an internal listserv for women in Google Brain —was more direct, accusing the company of "silencing marginalized voices" and calling its diversity efforts "a waste of time" (Simonite 2021). This email was sent to a group of female AI researchers that Gebru worked with in her department. The email was sent informally as Gebru voiced her frustrations to a team that she spent years working with.

While relaxing in an Airbnb in Austin the next evening, Gebru received a shocked message from a team member: "You resigned??" (Simonite 2021). Moments later, she found an email from Kacholia in her personal inbox. It rejected Gebru's offer and stated that her employment would end immediately, citing her internal email as "behavior inconsistent with Google's expectations of a manager." Her access to company systems was revoked, and Gebru tweeted that she had been fired. Google, however, maintained that she had resigned.

Retracting an academic paper would have been devastating for the researcher's academic career, background, and reputation. The retraction would call into question the validity of the research and suggests an association of misconduct (such as plagiarism and falsification) (Imafidon and McKie QC 2021). Gebru feared that her reputation would have been called into question and under fire.

To go forward, or not to go forward?

Ultimately, Gebru was faced with the decision of whether or not to go public about her circumstances, a concept known as whistleblowing. She evaluated several risks she faced if she went forward, including the heightened scrutiny of her team and their reputations faced. But what

does it truly mean to "blow the whistle" and why did Gebru view this act to be so consequential?

The legal definition of a whistleblower is important because it determines whether those that come forward can be provided legal protection for exposing the misconduct. A broad definition states that whistleblowing is the "disclosure by organizing members (former or current) of illegal, immoral, or illegitimate practices under the control of their employers, to persons or organizations that may be able to effect action" (Saade 2023). This definition requires that the whistleblower be a previous or current employee of the organization against which they are speaking. However, a competing definition removes the employment element, defining whistleblowing as the following:

Deliberate non-obligatory act of disclosure, which gets onto public record and is made by a person who has or had privileged access to data or information of an organization, about non-trivial illegality or other wrongdoing whether actual, suspected, or anticipated which implicates and is under the control of that organization, to an external entity having potential to rectify the wrongdoing. (Rehg, Miceli, and Van Scotter, 2008)

Often, the definition of whistleblowing includes an employment element. The Sarbanes-Oxley Act, a federal act that protects whistleblowers, includes applicable whistleblowers as "employees of public companies" (Spooner n.d.) which includes publicly traded companies and their subsidiaries, officers, contractors, subcontractors, and agents. Publicly traded companies are corporations whose shares are traded to the public on exchanges or over the counter (Banton 2025). Hence, under this Act, Dr. Gebru's decision to go forward would be protected since she was a Google employee, a company that was publicly traded. Some states even have whistleblower protection, like the California Whistleblower Protection Act, which prohibits retaliation against state employees who report waste, fraud, abuse of authority, violation of law, or threat to public health issues. The statute is also extended to include former employees (Saade 2023).

Dr. Gebru would be classified as a whistleblower because she would be exposing Google's mistreatment of her firing, raising her voice about her experiences of racism and sexism at work to the public (Perrigo 2022). This was privileged information that only she had access to as a Google employee. At the time of Gebru's firing, there were not many whistleblower protections present. In a Tweet days before Google fired her, Gebru asked whether anyone was working on regulation to protect AI ethics whistleblowers. Before she fired, she had voiced support for unionization as a means of protecting AI researchers; the Alphabet Workers Union cites Gebru's dismissal among the reasons it was formed. The union wrote a letter to state and national lawmakers that looked at potential policy outcomes of Gebru's firing—including unionization and changes to whistleblower protections laws. The analysis drew conversations with ethics, legal, and policy experts. UC Berkeley Center for Law and Technology co-director Sonia Katyal analyzed whistleblower protection laws in 2019, in the context of AI, and called them "totally insufficient. She argued that we should be concerned about a world where all the talented researchers like [Gebru] get hired at places like Google but are muzzled about speaking. When that happens, whistleblower protections become essential (Johnson 2021)."

The Evolution of Whistleblowing

Historically, when whistleblowers go public they are perceived extremely negatively, as an employee betraying their organization. Siri Nelson, the executive director of the National Whistleblowing Center, a nonprofit that provides legal assistance to whistleblowers and advocates protection laws, argues that "[s]ociety sort of looked at whistleblowers as rats or snitches." (US Securities and Exchange Commission n.d.). But more recently, perceptions about both the act of

whistleblowing and the whistleblowers themselves have changed, in large part because of federal protections like the Whistleblower Protection Act of 1989, the Dodd-Frank Act, the U.S. Securities and Exchange Commission SEC whistleblower program, and the Whistleblower Protection Enhancement Act of 2012 (US Securities and Exchange Commission n.d.). These laws changed the narrative, motivating people to come forward and help make public potential issues with the economy, public health, and a variety of other areas, in an effort to preemptively curtail further harm. Specifically, the Whistleblower Protection Act protects federal employes who disclose evidence of illegal activities, gross management, waste of funds, abuse of authority, or large/specific dangers to public health or safety within government agencies. The 2010 Dodd Frank Act expanded these whistleblower protections to the private sector and established monetary rewards for the money recovered by the government for individuals who voluntarily provided original information that led to enforcement of these laws in a court of law. The public's perception on whistleblowing has also changed drastically. A 2020 Marist Poll commissioned by Whistleblower Network News found that 86% of Americans strongly believe that whistleblowers who report corporate or government fraud should receive protection and 82% think that Congress should prioritize passing stronger laws to protect employees who report corporate fraud (Whistleblower Network News 2020).

Speaking Out in a "Boys' Club": Gender and the Cost of Truth-Telling

Whistleblowers often share one motivation: the desire to do what's right. Yet what doing the right thing means can differ depending on who you are, where you stand, and how the world treats you. Ethics are not universal—they're shaped by one's social position, gender, and sense of belonging within a workplace. Research has shown that women often approach ethical dilemmas through an ethics of care, focusing on relationships, empathy, and the impact of harm on others. Men, by contrast, tend to lean toward loyalty, justice, and rules (Valentine and Rittenburg 2007; Saade 2023).

For many women, especially those in male-dominated industries, the decision to blow the whistle isn't just about exposing wrongdoing—it's about reclaiming integrity in an environment that has long excluded them Women often feel shut out of the informal "boys' club" that defines corporate loyalty. Without access to these networks, women may feel less pressure to protect the system that marginalizes them. Men, on the other hand, are more likely to internalize loyalty to their peers and organizations, which can make speaking out feel like betrayal (Saade 2023).

Men often speak about regret and disloyalty. They see their actions as a break in team trust. Women, however, speak from a place of care. They expressed frustration at workplaces that made compassion impossible, describing whistleblowing as an act of moral alignment—bringing their work life closer to their personal values (Saade 2023).

This difference helps explain Dr. Timnit Gebru's decision to speak out. As one of the few Black women in the field of artificial intelligence, Gebru was used to being an outsider. When she raised concerns about bias in Google's large language models and the lack of diversity in the company's research culture, her questions weren't treated as collaboration—they were treated as defiance. In a space dominated by white and male leadership, her care for truth and justice didn't fit the culture of loyalty and silence. Her decision to challenge that silence was not a rejection of her field, but a deep act of care—for ethical science, for accountability, and for the people most harmed by biased technology.

Understanding Whistleblower Risks

Whistleblowing is an intensive process, involving legal, mental, and physical risks to one's well-being. For instance, for people involved in legal cases, the process is grueling and can result in individuals losing money and potentially incurring debt. Furthermore, whistleblowers can be legally restricted in what they can disclose to others, exacerbating anxieties and isolation. Related problems of post-traumatic stress disorder (PTSD) can lead to issues like clinical depression, anxiety, heart problems, hypertension, and related health concerns (van der Velden et al. 2019).

Jennifer Gibson, legal director of the Whistleblower Protection Program at the Signals Network, a nonprofit run by journalists and lawyers, describes that whistleblowers who go public represent a significant minority of concerned employees due to risks of being sent debilitating racist and sexist messages. Even when they report problems only internally, whistleblowers can face retaliation at work, including threats, intimidation, and gaslighting, demotions, firing, increased scrutiny of their daily responsibilities, ostracism, isolation, and gaslighting. The mental and emotional tax on truth-telling can strain relationships and families (Olumhense 2024).

Black women who report bad behavior at work jeopardize their professional careers, often weather more intense backlash when they report wrongdoing. This stems from the fact that women employees and employees of color are less likely to raise red flags than men or White employees as they are financially less able to take the risk, and they may have children they're caring for, they may have family they're caring for, and they may not be able to take the risk inherent in whistleblowing on a more sustained basis (Olumhense 2024). According to Siri Nelson, the biggest issue is thatpeople don't listen to Black women anyway and navigating that reality is something important to think about when you're considering blowing the whistle (Olumhense 2024). As one of the few Black women in tech, Dr. Timnit faced these microaggressions in her work environment and in response to her blowing the whistle.

When Black women do go public, the backlash often shifts the story away from the wrongdoing and toward their attitude, their tone, or their motives. Dr. Gebru experienced this pattern firsthand. After questioning the ethics of Google's research, Gebru was portrayed by Google as difficult, uncooperative, and angry—tropes historically weaponized against Black women in professional spaces. This was showcased in Google's retaliation in firing Gebru and the emails sent out to the public, explaining the situation from Google's viewpoint. The result was that the conversation moved away from her warnings about bias and toward debates about her professionalism.

Yet despite this, Gebru's act of truth-telling illuminated something larger: that whistleblowing, for women like her, is not simply about exposing corruption. It's about care in the face of indifference. It's about holding onto humanity in systems that reward silence. Her story reveals how gender, race, and ethics intertwine—how women who speak up often do so not because they feel safe, but because they know the cost of staying quiet is far greater.

Faced with a Complex Decision

Her decision to blow the whistle was difficult. Because Dr. Gebru was the first Black female research scientist at Google, there was an increased pressure to perform and complete her job in a significant manner. This concept is best explained by the previous analysis on the differences of the complexities of the whistleblowing decisions process amongst men and women.

Furthermore, from past stories of female whistleblowing – the risks of going forward were apparent. In an interview, Gebru reflected on Chelsey Glasso, who left Google in August 2019 due to pregnancy discrimination (Paul 2021). She had filed for a discriminatory lawsuit, a time-

consuming process. After Glasson left Google, she ended up working at Facebook. But, a few months into the new job, she was notified by Facebook's legal department that Google had subpoenaed her employee records – which included important information like her payroll tabs, performance evaluations and all complaint records she had filed while she worked there. In the discovery stage of her lawsuit, Glasson gave Google's legal team access to extremely private events in her life by force, including medical records and notes about her therapy sessions that had discussed her marriage and personal issues.

Glasson's experience forced Gebru to reflect on the exact risks she would face if she went forward with whistleblowing. It would not only open her up to a potential lawsuit, but also provide Google with the opportunity to violate her right to privacy. In an interview, Gebru stated that what happened to Glasson motivated her to not seek therapy in the months after her firing, in fear of these records being used for public fodder and brought into Google's control (Bhuiyan 2021).

But not only was Gebru in fear of the privacy risks she faced, she worried about the public response that Google would make. In the years prior to her firing, she already had difficulty in instilling change within Google's culture. As the first Black research scientist at Google, she described discussions about diversity initiatives to be extremely frustrating (Newton 2020). This was after attempts to connect with management about issues Gebru was facing within the company's culture that were shut down almost immediately. For example, Gebru had written millions of documents about diversity initiatives about racial literacy and machine learning fairness initiatives for the importance of women retention – but when she wanted to meet with management to go over these documents and implement them in future projects – it was met with extreme resistance. Hence, without the proper oversight and addressing of policies, Dr. Timnit did not want to pursue the issue any further. She compared it to a situation where "someone is shooting at you with a gun, and you're screaming. Instead of trying to stop the person from shooting you, they're trying to stop you from screaming" (Hao 2020)

.

Conclusions

Timnit Gebru's departure from Google illustrates a broader tension between corporate commitments to ethical artificial intelligence and the realities of power and voice within large technology firms. Although Google has publicly pledged to eliminate bias, decision-making authority remains concentrated among leadership that has historically marginalized minority perspectives, often silencing internal critics who raise ethical concerns (Metz and Wakabayashi 2020). Gebru's firing exposed how corporate support for "AI for social good" frequently extends only as far as it benefits public relations, underscoring the risks faced by whistleblowers in the technology sector. Her situation brings to light a few questions: Do female whistleblowers face a harsher standard compared to men whistleblowers? How does race play a role in evaluating female leaders in an dynamic industry, like artificial intelligence? Would Timnit Gebru have received the same backlash had she not been outspoken about her racist experiences at Google? Timnit Gebru brought light to an important conversation about inequitable artificial intelligence programs and led discussions on how industry leaders should proceed with warnings of AI's dangers. Timnit's background provided her a unique lens in understanding potential dangers of AI. How else should senior leadership in prominent technology companies assess feedback from their own employees? How can we prevent artificial intelligence from harming marginalized communities?

Epilogue

Timnit's Eventual Decision

Eventually, Timnit decided to go forward about her experiences. But, when Google unceremoniously ousted Dr. Timnit Gebru in 2020, she felt targeted. As a Black woman, Dr. Timnit Gebru faced doubt and intense scrutiny from her managers within Google. Gebru described how after the disclosure, she was victim to a lot of swearing, stalking, harassment, death threats, daily emails with the N-word, and other slurs, on top of losing her job. The following weeks after her firing, she decided to publicly release the email she had sent her department (See Appendix A).

Yet, she reminded herself that many of those who spoke before her had disappeared, and no matter how brave or strong they were, they were just not living in a time that was supportive and where they would be believed. She describes that there were two words that kept her going: collective organizing. She had spent a long time building a network, and whether that was at Google or outside, she had a strategy to build a network with collective support or power. Whistleblowing is often an extremely solitary and isolating activity. However, Timnit's experience was unique, given her community's dedication to supporting her through public backlash (Berkeley 2022).

Google's Response to the Paper

In a public explanation on Google Docs and email to Google's staff (see Appendix B), Jeff Dean explained the reasoning behind Google's request. Jeff explains the circumstances around Dr. Timnit's submission of the paper. Google has approved dozens of papers that Dr. Timnit and other Google authors have submitted or authored and subsequently published. However, these papers undergo several changes through the review process, and some are even deemed unsuitable for publication. Jeff Dean claimed that the particular paper published was only submitted with one day's notice before the deadline and that it typically takes two weeks to review a paper. Instead of waiting for reviewers' feedback, it was approved automatically for submission and published. Furthermore, the paper ignored relevant research: because it talked about environmental impacts of large models, disregarding models with greater efficiency. Similarly, the paper raised concerns of biases within language models but did not consider recent research mitigating these issues (Dean 2020).

In response to criticisms made on Google's devotion to ethics in their research and review process, Jeff explains the process in depth. The process involves more than just a single approver and a small team of immediate researchers. It also includes a wide range of researchers, social scientists, ethicists, policy & privacy advisors, and human rights specialists across all disciplines. For every 1000 papers that are successfully published, there are several that do not end up in publication. Dr. Gebru's paper had "important gaps" that prevented Google from putting their name on it. For a successful publication, papers are given extensive feedback that make them stronger than initially submitted. Jeff lists several papers that challenge Gebru's claims on the risk of AI models – some that engage in the diversity initiatives that Timnit criticizes within the technological world. Some topics covered include measuring and reducing gender correlations in pre-trained models, evading Deep Fake-Image Detectors with White and Black-Box Attacks, and what AI means for small farmers.

Why Timnit's experience matters in the world of Whistleblowing and Technology

Google has repeatedly committed to eliminating bias in its systems, but the trouble is within

the hiring system (Metz and Wakabayashi 2020). Dr. Gebru said that most of the ultimate decisions are made by men, who are "not only failing to prioritize hiring more people in minority communities, but they are also quashing their voices." Dr. Gebru's departure has reflected a larger problem within the industry, showing how some large tech companies only support ethics and fairness and other artificial intelligence for social good causes, as long as their positive public relations impact outweighs the public scrutiny that it brings (Metz and Wakabayashi 2020). Her situation drew massive attention to the importance of whistleblowing, specifically within large technology companies like Google.

After Gebru was fired, she received the backing of other company members and letters from Congress. She regrouped and launched the Distributed Artificial Intelligence Research Institute (DAIR), which continued research about the dangers of using AI, that was frowned upon. DAIR works with computer scientists and engineers but goes a step further by working with individuals who have directly been impacted by AI's harmful side. When she created DAIR, Dr. Gebru wanted to make sure that the institute's work was shaped not only by computer scientists and engineers, but also by people who have lived experience in researching harms that have perpetuated their lives through technology (Dorisca 2025).

Prior to this incident, Dr. Timnit had already co-founded the Black in AI event at the Neural Information Processing Systems conference in 2017, with the motivation to educate and address how biases get into AI systems and how to counteract them. She explained that diversity in AI is important not only within data sets, but also in researchers where one needs a social sense of where things are. Accordingly, we are in a diverse crisis for AI, and this is an urgent issue. To do this, there are several approaches. "One includes diversifying a data set and having many different annotations about the data set, like gender and age. Then, once you train a model, one can test it out and see how well it is by different subgroups" (Snow 2018).

Dr. Gebru continues to serve on the board of Addis Coder, a non-profit dedicated to teaching Ethiopian and Jamaican high school students, receiving several accolades—including Nature's top ten people who helped shape science and one of TIME 100's influential people. In the process of doing so, she is also writing a memoir and manifesto, "The View from Somewhere," that argues for a technological future that serves communities, instead of one that focuses on warfare or surveillance (DAIR 2025). Today, her advocacy has brought light and emphasis on supporting marginalized communities who are directly affected by artificial intelligence systems.

Dr. Gebru's story teaches the importance of having women of color voices in a technologically innovative company that can bring about harm on communities who don't have an impact or voice on the research done on their lives within the company.

References

Allison, Simon. 2022. "Timnit Gebru and the Fight to Make Artificial Intelligence Work for Africa." *The Mail & Guardian*. https://mg.co.za/africa/2022-06-09-timnit-gebru-and-the-fight-to-make- artificial-intelligence-work-for-Africa/ (Accessed March 26, 2025).

Allyn, Bobby. 2020. "Ousted Black Google Researcher: 'They Wanted To Have My Presence, But Not Me Exactly.'" *NPR*. https://www.npr.org/2020/12/17/947719354/ousted-black-google-researcher-they- wanted-to-have-my-presence-but-not-me-exactly (Accessed April 29, 2025).

Berkeley, UC. 2022. "Timnit Gebru to UC Berkeley Graduates: 'Work Collectively for a Better Future for All' | UC Berkeley School of Information." https://www.ischool.berkeley.edu/news/2022/timnit- Gebru-uc-Berkeley-graduates-work-collectively-better-future-all (Accessed April 3, 2025).

Brown, Sara. 2021. "Ex-Google Researcher: AI Workers Need Whistleblower Protection | MIT Sloan." *MIT Management Sloan School*. https://mitsloan.mit.edu/ideas-made-to-matter/ex-google-researcher-ai- workers-need-whistleblower-protection (Accessed March 26, 2025).

Dean, Jeff. 2020. "About Google's Approach to Research Publication." *Google Docs*. https://docs.google.com/document/d/1f2kYWDXwhzYnq8ebVtuk9CqQqz7ScqxhSIxeYGrWjK0/edi t?usp=sharing&usp=embed_facebook (Accessed March 26, 2025).

Dorisca, Samantha. 2025. "After Being Fired From Google, Timnit Gebru Launched An AI Research Institute That Is Not Bound To BigTech's Influence." *AfroTech*. https://afrotech.com/interview- timnit-gebru (Accessed April 2, 2025).

Hao, Karen. 2020. "We Read the Paper That Forced Timnit Gebru out of Google. Here's What It Says." *MIT Technology Review*. https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics- research-paper-forced-out-timnit-gebru/ (Accessed March 26, 2025).

Harris, John. 2023. "'There Was All Sorts of Toxic Behaviour': Timnit Gebru on Her Sacking by Google, AI's Dangers and Big Tech's Biases." *The Guardian*. https://www.theguardian.com/lifeandstyle/2023/may/22/there-was-all-sorts-of-toxic-behaviour-timnit-gebru-on-her-sacking-by-google-ais-dangers-and-big-techs-biases (Accessed December 18, 2025).

Harris, John. 2023. "'There Was All Sorts of Toxic Behaviour': Timnit Gebru on Her Sacking by Google, AI's Dangers and Big Tech's Biases." *The Guardian*. https://www.theguardian.com/lifeandstyle/2023/may/22/there-was-all-sorts-of-toxic-behaviour- timnit-gebru-on-her-sacking-by-google-ais-dangers-and-big-techs-biases (Accessed March 25, 2025).

Hunt, Linda. 2016. 9 *An Examination of the Role Women Whistleblowers*. International Business Research. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2871449. (Accessed December 18, 2025).

Imafidon, Anne-Marie, and Suzanne McKie QC. 2021. "Timnit Gebru, Google, and Institutional Discrimination in AI: Lessons for 2021 - IFOW." https://www.ifow.org/news-articles/timnit-gebru- google-and-institutional-discrimination-in-ai-lessons-learned-for-2021 (Accessed April 1, 2025).

Johnson, Khari. 2021. "Google Employee Group Urges Congress to Strengthen Whistleblower Protections for AI Researchers." *VentureBeat*. https://venturebeat.com/ai/google-employee-group-urges- congress-to-strengthen-whistleblower-protections-for-ai-researchers/ (Accessed April 3, 2025).

Kenny, Kate, and Marianna Fotaki. 2023. "The Costs and Labour of Whistleblowing: Bodily Vulnerability and Post-Disclosure Survival." *Journal of Business Ethics* 182(2): 341–64. doi:10.1007/s10551-021- 05012-x. (Accessed March 20, 2025).

Kenyon, Tilly. 2021. "DIAR: An AI Research Institute Created by Ex-Google Employee." https://aimagazine.com/ai-strategy/diar-ai-research-institute-created-ex-google-employee (Accessed April 2, 2025).

McShane, Julianne. 2022. "Timnit Gebru Is Part of a Wave of Black Women Working to Change AI." *NBC News*. https://www.nbcnews.com/news/nbcblk/timnit-gebru-part-wave-black-women-working- change-ai-rcna13339 (Accessed April 2, 2025).

Metz, Cade, and Daisuke Wakabayashi. 2020. "Google Researcher Says She Was Fired Over Paper Highlighting Bias in A.I." *The New York Times*. https://www.nytimes.com/2020/12/03/technology/google-researcher-timnit-gebru.html (Accessed March 26, 2025).

Newton, Casey. 2020. "The Withering Email That Got an Ethical AI Researcher Fired at Google." (Accessed December 18, 2025).

O'Leary, Lizzie. 2020. "A Black A.I. Ethicist on Her Experiences at Google—and Her Controversial Departure." *Slate*. https://slate.com/technology/2020/12/google-timnit-gebru-paper-black- employees.html (Accessed March 26, 2025).

Olumhense, Ese. 2024. "Whistleblowing While Black: How Truth-Telling Changes the Careers of Black Women in Tech." *The 19th*. https://19thnews.org/2024/03/black-women-tech-whistleblowers/ (Accessed April 3, 2025).

Perrigo, Billy. 2022. "Timnit Gebru on Not Waiting for Big Tech to Fix AI | TIME." https://time.com/6132399/timnit-gebru-ai-google/ (March 25, 2025).

Perrigo, Billy. 2022. "Why Timnit Gebru Isn't Waiting for Big Tech to Fix AI's Problems." *TIME*. https://time.com/6132399/timnit-gebru-ai-google/ (Accessed December 18, 2025).

Platformer. https://www.platformer.news/the-withering-email-that-got-an-ethical/ (Accessed March 29, 2025).

Rocha, Gene Da. 2023. "A Brief History on Timnit Gebru — A.I. Researcher." *Medium*. https://medium.com/@genedarocha/a-brief-history-on-timnit-gebru-a-i-researcher-cae368d45b49 (Accessed March 24, 2025).

Saade, Mary. 2023. "Women & Whistleblowing." https://repository.uclawsf.edu/hwlj/vol34/iss1/12 (Accessed March 25, 2025).

Snow, Jackie. 2018. "'We're in a Diversity Crisis': Cofounder of Black in AI on What's Poisoning Algorithms in Our Lives." *MIT Technology Review*. https://www.technologyreview.com/2018/02/14/145462/were-in-a-diversity-crisis-black-in-ais- founder-on-whats-poisoning-the-algorithms-in-our/ (Accessed March 24, 2025).

Spooner, Lyn. "Sarbanes-Oxley Act | Sarbanes-Oxley Compliance Professionals Association (SOXCPA)." https://www.sarbanes-oxley-act.com/ (Accessed April 29, 2025).

Strubell, Emma. "Carbon Footprint of AI and Deep Learning." https://www.learningtree.com/blog/carbon- footprint-ai-deep-learning/ (Accessed May 2, 2025).

Team, AI4ALL. 2018. "Role Models in AI: Timnit Gebru." *AI4ALL*. https://medium.com/ai4allorg/role-models-in-ai-timnit-gebru-e4ad53b01366 (Accessed December 18, 2025).

Thomas, Rachel. 2020. "The Far-Reaching Impact of Dr. Timnit Gebru." *The Gradient*. https://thegradient.pub/the-far-reaching-impacts-of-timnit-gebru/ (Accessed March 24, 2025).

US Securities and Exchange Commission. "SEC.Gov | Whistleblower Protections." https://www.sec.gov/enforcement-litigation/whistleblower-program/whistleblower-protections (Accessed April 29, 2025).

Valentine, Sean, and Terri L. Rittenburg. 2007. "The Ethical Decision Making of Men and Women Executives in International Business Situations." 71 J. Bus. Ethics, 125. doi:10.1007/s10551-006- 9129-y (Accessed December 18, 2025).

Walsh, Dylan. 2022. "Timnit Gebru: Ethical AI Requires Institutional and Structural Change." *Stanford Human-Centered Artificial Intelligence*. https://hai.stanford.edu/news/timnit-gebru-ethical-ai-requires-institutional-and-structural-change (Accessed December 18, 2025).

Whistleblower Network News, Marist Poll. 2020. "Marist Survey Results." *Whistleblower Network News*. https://whistleblowersblog.org/whistleblower-news-network-survey (Accessed April 24, 2025).

Appendix A: Timnit's Email to Google Brain Women and Allies

On December 2nd, 2020, Dr. Timnit wrote an email to Google's Brain Women and Allies. After being fired from Google, she decided to write to her co-workers explaining the circumstances around her work in diversity, equity, and inclusion. She expressed her frustrations in getting change within Google's culture. She provided her perspective on what truly happened with her published paper:

Hi friends,

I had stopped writing here as you may know, after all the micro and macro aggressions and harassments I received after posting my stories here (and then of course it started being moderated).

Recently however, I was contributing to a document that Katherine and Daphne were writing where they were dismayed by the fact that after all this talk, this org seems to have hired 14% or so women this year. Samy has hired 39% from what I understand but he has zero incentive to do this.

What I want to say is stop writing your documents because it doesn't make a difference. The DEI OKRs that we don't know where they come from (and are never met anyways), the random discussions, the "we need more mentorship" rather than "we need to stop the toxic environments that hinder us from progressing" the constant fighting and education at your cost, they don't matter. Because there is zero accountability. There is no incentive to hire 39% women: your life gets worse when you start advocating for underrepresented people, you start making the other leaders upset when they don't want to give you good ratings during calibration. There is no way more documents or more conversations will achieve anything. We just had a Black research all hands with such an emotional show of exasperation. Do you know what happened since? Silencing in the most fundamental way possible.

Have you ever heard of someone getting "feedback" on a paper through a privileged and confidential document to HR? Does that sound like a standard procedure to you, or does it just happen to people like me who are constantly dehumanized?

Imagine this: You've sent a paper for feedback to 30+ researchers, you're awaiting feedback from PR & Policy who you gave a heads up before you even wrote the work saying "we're thinking of doing this," working on a revision plan figuring out how to address different feedback from people, haven't heard from PR & Policy besides them asking you for updates (in 2 months). A week before you go out on vacation, you see a meeting pop up at 4:30pm PST on your calendar (this popped up at around 2pm). No one would tell you what the meeting was about in advance. Then in that meeting your manager's manager tells you "It has been decided" that you need to retract this paper by next week, Nov. 27, the week when almost everyone would be out (and a date which has nothing to do with the conference process). You are not worth having any conversation about this, since you are not someone whose humanity (let alone expertise recognized by journalists, governments, scientists, civic organizations such as the electronic frontiers foundation etc.) is acknowledged or valued in this company.

Then, you ask for more information. What specific feedback exists? Who is it coming from? Why now? Why not before? Can you go back and forth with anyone? Can you understand what exactly is problematic and what can be changed?

And you are told after a while that your manager can read you a privileged and confidential

document and you're not supposed to even know who contributed to this document, who wrote this feedback, what process was followed or anything. You write a detailed document discussing whatever pieces of feedback you can find, asking for questions and clarifications, and it is completely ignored. And you met with, once again, an order to retract the paper with no engagement whatsoever.

Then you try to engage in a conversation about how this is not acceptable, and people start doing the opposite of any sort of self-reflection—trying to find scapegoats to blame.

Silencing marginalized voices like this is the opposite of the NAUWU principles which we discussed. And doing this in the context of "responsible AI" adds so much salt to the wounds. I understand that the only things that mean anything at Google are levels, I've seen how my expertise has been completely dismissed. But now there's an additional layer saying any privileged person can decide that they don't want your paper out with zero conversation. So you're blocked from adding your voice to the research community—your work which you do on top of the other marginalization you face here.

I'm always amazed at how people can continue to do things after things like this and then turn around and ask me for some sort of extra DEI work or input. This happened to me last year. I was in the middle of a potential lawsuit for which Kat Herller and I hired feminist lawyers who threatened to sue Google (which is when they backed off--before that Google lawyers were prepared to throw us under the bus, and our leaders were following as instructed) and the next day I get some random "impact award." Pure gaslighting.

So, if you would like to change things, I suggest focusing on leadership accountability and thinking through what types of pressure can also be applied from the outside. For instance, I believe that the Congressional Black Caucus is the entity that started forcing tech companies to report their diversity numbers. Writing more documents and saying things over and over again will tire you out but no one will listen.

- Timnit

Appendix B: Jeff Dean's Email to Google

On December 3rd, 2020, Google AI Chief Jeff Dean gave a response to Timnit Gebru's calls of action:

Hi everyone,

I'm sure many of you have seen that Timnit Gebru is no longer working at Google. This is a difficult moment, especially given the important research topics she was involved in, and how deeply we care about responsible AI research as an org and as a company.

Because there's 3been a lot of speculation and misunderstanding on social media, I wanted to share more context about how this came to pass, and assure you we're here to support you as you continue the research you're all engaged in.

Timnit co-authored a paper with four fellow Googlers as well as some external collaborators that needed to go through our review process (as is the case with all externally submitted papers). We've approved dozens of papers that Timnit and/or the other Googlers have authored and then published, but as you know, papers often require changes during the internal review process (or are even deemed unsuitable for submission). Unfortunately, this particular paper was only shared with a day's notice before its deadline — we require two weeks for this sort of review—and then instead of awaiting reviewer feedback, it was approved for submission and submitted.

A cross functional team then reviewed the paper as part of our regular process and the authors were informed that it didn't meet our bar for publication and were given feedback about why. It ignored too much relevant research — for example, it talked about the environmental impact of large models, but disregarded subsequent research showing much greater efficiencies. Similarly, it raised concerns about bias in language models, but didn't take into account recent research to mitigate these issues. We acknowledge that the authors were extremely disappointed with the decision that Megan and I ultimately made, especially as they'd already submitted the paper.

Timnit responded with an email requiring that a number of conditions be met in order for her to continue working at Google, including revealing the identities of every person who Megan and I had spoken to and consulted as part of the review of the paper and the exact feedback. Timnit wrote that if we didn't meet these demands, she would leave Google and work on an end date. We accept and respect her decision to resign from Google.

Given Timnit's role as a respected researcher and a manager in our Ethical AI team, I feel badly that Timnit has gotten to a place where she feels this way about the work we're doing. I also feel badly that hundreds of you received an email just this week from Timnit telling you to stop work on critical DEI programs. Please don't. I understand the frustration about the pace of progress, but we have important work ahead and we need to keep at it.

I know we all genuinely share Timnit's passion to make AI more equitable and inclusive. No doubt, wherever she goes after Google, she'll do great work and I look forward to reading her papers and seeing what she accomplishes.

Thank you for reading and for all the important work you continue to do.

- Jeff