When diving deep into operationalizing the ethical development of artificial intelligence, one immediately runs into the "fractal problem". This may also be called the "infinite onion" problem.[1] That is, each problem within development that you pinpoint expands into a vast universe of new complex problems. It can be hard to make any measurable progress as you run in circles among different competing issues, but one of the paths forward is to pause at a specific point and detail what you see there.
Here I list some of the complex points that I see at play in the firing of Dr. Timnit Gebru, and why it will remain forever after a really, really, really terrible decision.

The Punchline. The firing of Dr. Timnit Gebru is not okay, and the way it was done is not okay. It appears to stem from the same lack of foresight that is at the core of modern technology,[2] and so itself serves as an example of the problem. The firing seems to have been fueled by the same underpinnings of racism and sexism that our AI systems, when in the wrong hands, tend to soak up. How Dr. Gebru was fired is not okay, what was said about it is not okay, and the environment leading up to it was -- and is -- not okay. Every moment where Jeff Dean and Megan Kacholia do not take responsibility for their actions is another moment where the company as a whole stands by silently as if to intentionally send the horrifying message that Dr. Gebru deserves to be treated this way. Treated as if she were inferior to her peers. Caricatured as irrational (and worse). Her research writing publicly defined as below the bar. Her scholarship publicly declared to be insufficient. For the record: Dr. Gebru has been treated completely inappropriately, with intense disrespect, and she deserves an apology.

Background: Ethical AI Approach to Developing Technology. I had come from Microsoft to Google to spearhead a new approach to research, where we take a step back for the "bigger picture". Research questions could be grounded in human values, the inclusion of diverse experiences, and learning from multiple time points and social movements. In this approach, both learning from the past and *foresight* are prioritized.[3] The idea is that, to define AI research *now*, we must look to where we want to be in the future, working backwards from ideal futures to this moment, right now, in order to figure out what to work on today. This gives rise to an approach that can only function well with inclusion of diverse experiences.[4] And such inclusion can only function well if the individuals ***belong***, and are treated that way.

This is a fundamentally different approach than the common operating paradigm where the goal is to do "something novel" or to improve a given task. The forward-looking approach I'm fascinated by ignores these tasks altogether and instead asks, "what could AI do to bring about a better society?"

---

[1] Which, to be fair, is a better band name.
[2] And diversity and inclusion efforts.
[3] Learning from the past also helps with foresight.
[4] If you don't understand why, don't worry, we know what we're doing and have explained in multiple presentations.

An approach grounded in human values, long-term beneficial outcomes, social patterns, diversity, and inclusion can be broadly referred to as “Ethical AI”.

The Ethical AI Team. If this is your cup of tea, then in this frame of mind, it is not hard to think through how AI could, ideally, provide ways for equal access to opportunity and beneficial outcomes. It is also not hard to see how it could massively, massively mess that up. When you can ground your research thinking in both foresight and an understanding of society,[5] then the research questions to currently focus on fall out from there. For example, it becomes clear that in order for systems to be used in the best ways possible in the future, then today there needs to be research on mechanisms to report how well systems work and holistic development approaches to mitigate runaway feedback loops.

After spending two not-so-awesome years (but working with mostly awesome people) laying the groundwork for a foundation of “ML Fairness”, I found myself leading a tiny team that believed in these fundamental ideas of evolving AI with long-term thinking. We sought to extricate ourselves from negative interpersonal politics arising from our oddball approach to research, and tried to focus on what really mattered: the future, AI, and society. Two years ago, I thought we were in a really good place, and I was ecstatic that Dr. Gebru accepted my invitation to join us as a team co-lead.

Co-Leading. I wanted to co-lead with Dr. Gebru because I thought that the two of us together could foster an incredible team. We were already aligned on the idea of developing AI in a new way, with diversity and inclusion at its core, and informed by an understanding of human beliefs, values, and how these interplay with technology. Dr. Gebru is also an outstanding leader in some of the ways I am the weakest. She is a visionary leader in AI development, while working non-stop on inclusion. Her work includes statistical research on the discrepancy between academic models and the real-world, data-driven work demonstrating that socioeconomic

[5] Not a task for the weak-willed.

attributes of different regions can be inferred from publicly available images, including estimating per-capita carbon emissions, work providing evidence of different race[6] and gender error patterns in AI systems. As a founder of Black in AI, she has also single-handedly increased the number of Black people participating in major AI conferences globally by orders of magnitude and has vastly expanded the network of Black people working in AI.[7] Dr. Gebru's leadership style is empathetic and driven, and she can identify -- and often fix -- unjust treatment of the people she leads.[8] She has led the way forward on massive-scale nuanced and complicated problems, everything from strengthening Google's research lab in Accra, Ghana to helping Black researchers whose visas are being disproportionately denied to hosting a major AI conference in her hometown of Addis Ababa, Ethiopia while navigating the complexities of LGBTQ+ safety. It is clear[9] that at **this** moment in the story of AI, the path to an intelligence that will not harm those *most at risk of being harmed* requires Dr. Gebru's abilities, skill, and deep knowledge.

With Dr. Gebru, our team flourished and grew. Dr. Gebru and I had just been promoted to "Staff" Research Scientist, which is meaningful in the world of STEM, with the sort of honor associated with becoming tenured. We had thought it meant a certain amount of job security.

What Happened. Different job roles have different job incentives. Making progress in a context of multiple conflicting incentives is hard enough, but in addition, incentives are weighted by your "level" in a hierarchy. The higher a person's "level", the more weight there is for their incentives. This can cause a very top-heavy drag on ideas that are obviously very dumb and hard to understand.[10] Such as firing Dr. Timnit Gebru and calling it a resignation.

What Now. Dr. Gebru refused to subjugate herself to a system[11] requiring her to belittle her integrity as a researcher and degrade herself below her fellow researchers.[12] Within the next year, let those of us in positions of privilege and power come to terms with the discomfort[13] of being part of an unjust system that devalued one of the world's leading scientists, and keep something like this from ever happening again. Dr. Alex Hanna and Meredith Whittaker outlined some of how this can happen and I look forward to the fundamental shift in excessively poorly focused power.

---

[6] Actually, it's skin "type", which is race-correlated (but pause on that for now; it's another fractal path).
[7] "Wikipedia famous".
[8] Including the people who reach out to her.
[9] ...to those who can wrap their head around this stuff
[10] Sometimes they can be understood if you can wrap your head around intensely focusing on the "here and now", the short-term, and the limited input of a small set of relatively homogenous collaborators.
[11] Exactly 65 years earlier, Rosa Parks refused to subjugate herself to a system that said she must be treated with less respect than her fellow passengers. Within a year, the unjust process that she refused to participate in was ruled unconstitutional.
[12] None of whom were required to blindly obey mandates from a "single-blind" review process known to be a vehicle for sex-based bias.
[13] Don't worry. You can do it!