Opening Thoughts of Margaret Mitchell
Chief Ethics Scientist and Researcher
Hugging Face

For the AI and the Future Democratic Caucus
discussion with leading experts on the risks of Artificial Intelligence (AI)
July 26, 2023

Trying to foresee and predict all possible harms from an AI system can feel a bit like trying to catch fish in the sea with bare hands. How can you predict all the different ways a system can be problematic?

- As you start to understand the different areas of harm, and how solutions on the one hand can create problems on the other, it can begin to feel overwhelming.
- A key question here today is how to focus attention on the negative consequences of technology in a way that is well-informed, nuanced, and coherent. I hope to help with this today.

I've worked in what is here being called "artificial intelligence", in academic, non-profit, and corporate settings, for almost 20 years (including studying under Dr. Emily Bender here today).

- My research work includes detailed publications on 3 things that are critical to responsible AI development:
    - documentation,
    - governance,
    - and auditing.
- I'm currently working at Hugging Face, a platform for open-source AI, and for the past 10 years, I've worked within both more open and more closed technology companies, advancing AI while also guiding people in how to think through different outcomes of technology: What we can do to bring about the outcomes we want, while not bringing about those we don't?

Let me first get to the punchline. While the promises of technology can be incredibly exciting, the solutions for safe and ethical AI are generally not exciting – they're rather dry and boring. Which is part of why only a small minority of people, including those sitting here today, have seriously engaged in working on them. So, as an expert with years of experience in this area, here is my perspective how to develop safe and ethical AI:

- Clear articulation of goals before development begins
    - This unpacks into a lot that I hope to discuss;

        - includes goals for downstream users as well as those involved in creating the system in the first place, and who *should* be involved when the goal is to benefit humanity
- Rigorous curation and measurement of data that is input to training
    - This helps to ensure that we have control over what the system learns.
- Diverse evaluation informed by foreseeable use in context
    - This helps to understand the situations and subpopulations the system will work well for, as well as those they won't work well for
- And robust documentation throughout the end-to-end machine learning development pipeline (which I can sketch out for you if useful), including considerations and decisions, that is made transparent for others to view.
    - This incentivizes good practices – if people are looking at your work, you're going to make sure the work is good
    - And documentation protocol drives people to think about things that might not otherwise occur to them.

Now let's talk through the prompt for this meeting.

- In order to anticipate **risks**, the critical skill is **foresight**.
    - Some people are better than this than others (and this fact is usually overlooked).
    - Some companies will have you believe it's not possible to do this well. It is.
- In order to identify current and future **harms**, and their seriousness, the critical ingredient is **life experience**.
- Those people who have been subject to harms from society, and from technology directly – people with characteristics that are marginalized – are not unsurprisingly often the best equipped to recognize the risk of those harms emerging. In this way, **systemic discrimination in tech directly results in problematic technology.**

One example of harms and risks includes the **risk** of people believing the output of systems, which can realize itself in different **harms** in different contexts.

- This includes life-critical situations (medicine, legal domains)
- This includes manipulation of people to do problematic things, which can be personalized, targetted, and deployed to billions.
    - For example, persuasion towards extremism and violent acts; influencing minors towards sexual acts; and scamming older people less familiar with technology.

  (This also includes convincing people that what is real is fake and what's fake is real)

A related example includes the **risk** of people being subconsciously influenced by problematic viewpoints that these systems implicitly encode, which include views that are discriminatory to

marginalized populations – for example, racism and sexism. This creates the **harms** of inequity and inequality.

- These views can be incredibly subtle, for example, in text, this can be the order that results are presented in, where white men lead, or the over-association of some concepts with different identities, such as over-associating women and "smile".
- In images, this can be the positioning and spatial relations of those depicted, where for example white men are more likely to be centered;
- These views can be subtly influential in how we think about & perceive the world, enforcing stereotypes, and further propagating inequality.

One approach to address these issues, and to creating values-informed AI development, is to turn harm-based focus on its head, to focus on *rights* that people should not have broken or disrupted.

- I find it a bit easier to figure out how to help shape the evolution of AI in a way that's more aligned with peoples' values and goals by making **those values the end goal** of the system and figuring out what needs to be done to get there.
    - This gets into the distinction between "general purpose" and "task focused" systems, which is roughly the AGI vs ANI distinction.
    - What you find is that it's task-focused models, not "general" models, that are most appropriate for AI to be aligned to human values.
        - And I have a lot to unpack about what "general" really means.
- By centering people, what they need, and their rights;
    - And figuring out how to work towards those things
- Instead of centering technology
    - And trying to figure out how it maps to values post-hoc

In summary, my advice for addressing harms and risks is to center on people and their values from the start. The current direction of ill-specified, so called "general" models, with values applied post-hoc, is exactly the wrong direction for more beneficial AI.