Statements for AI Insight Forum: High Impact AI
11/01/2023
Margaret Mitchell

Introduction

Good afternoon, Leader Schumer, Senator Rounds, Senator Heinrich, and Senator Young. Thank you for the opportunity to speak with you today about high impact artificial intelligence. My name is Margaret Mitchell, and I'm an AI researcher and Chief Ethics Scientist at the AI start-up Hugging Face. I have been working in what is now referred to as "AI" for the past 18 years, in academic research labs and private technology companies. I have worked on operationalizing ethics-informed artificial intelligence throughout the tech industry, including in two "Big Tech" companies, Microsoft and Google, and in the largest open community platform for starts-up and academics, Hugging Face. My ethics work has focused on how to identify implicit values and biases throughout AI development and how these affect people impacted by AI systems. My work has shown me that it is critical to assess the impact of AI systems with respect to **rights**, including human rights, civil rights, and cultural rights. It has also shown me that rigorous, robust documentation, with clear articulation of the tasks a system may be used for, how it may be used in different contexts, and evaluation of how well it can be expected to perform across these different contexts, is key to advancing AI aligned with human values while minimizing potential harms to individuals.

To understand which AI applications are "high impact" and how to regulate them, we need to center the people impacted, rather than the technology itself. This is an approach that examines the effects of AI on the rights of the people behind its development and the people subject to its outputs. By defining relevant stakeholder groups and subpopulations (such as gender, ethnicity, or age subgroups), we can derive the benefits, harms, and risks to different people in different AI contexts, identifying the potential for positive and negative impact. Stakeholder groups include the following, which is non-exhaustive and not mutually exclusive:

- **Data creators and data subjects**, including those producing "raw" data (such as artists), those annotating it (such as crowdworkers), and the people represented in the data
- **AI developers**, which may be individual engineers or larger organizations (such as tech companies)
- **AI deployers**, who leverage the technology for different applications (such as companies and government agencies)
- **AI users**, who interact with the technology made available by deployers (such as people in education, healthcare, and finance)
- **AI-affected**, who may have AI technology applied to them, whether or not they chose to (such as in surveillance)

Data creators make up a large stakeholder group for AI technology, and thus should be given special attention when evaluating what constitutes "high impact" applications. They are the people who create the texts, audio recordings, videos, and images that are used in AI training. Current issues include whether they have rights to **consent**, **credit**, and **compensation**. Ideally, when someone produces content and makes it available in a form where an AI developer may use it – such as by posting it on the internet – they should be able to provide guidance on what is used and how. At a high level, this means they should be able to specify details such as which of their shared content can be used for AI training, who can see it, how it can be processed, and the kinds of AI systems it can be used for. Such specifications are types of *informed consent*. Data creators should ideally also be able to define how the use of their work in an AI system connects back to them, such as via *credit*, where they are acknowledged alongside AI output, and via *compensation*, where they are paid for the usage of their work in AI training, the demand for AI outputs influenced by their work, or both. Here, a high negative impact would be loss of rights for fair compensation and opportunity due to their work being cycled through an AI system that replaces them. For *data annotators*, current issues within the U.S. include remuneration well below minimum wage (e.g., Hornuf & Vranker, 2022) through companies that can well afford to fairly compensate their workers; and internationally, include digital colonialism, where people from countries with a lower cost of living are given the opportunity to annotate data for compensation – at the expense of rights to healthcare, work-life balance, and enriching the nation where they work (Mohamed et al, 2020; Hao, 2022; Perrigo, 2023).

AI developers and deployers like me constitute another important stakeholder group, which has been well-represented in recent regulatory discussions. For us, there is a clear net benefit in advancing AI: As demand for AI increases, so too do our job opportunities, our salaries, and our privilege to inform legislation. I am thankful for this privilege and honor, and so hope to use it, as best I can in my capacity as an ethics practitioner, machine learning scientist, and researcher, to help connect the dots between how technology companies operate and the impact of the AI we create across subpopulations and other stakeholder groups.

The majority of current discussion in regulatory circles on the impact of AI focuses on the last groups that I've listed, **AI users and AI-affected**, in sectors such as healthcare, medicine, education, and finance. In these domains, AI has the potential to uncover new insights and advance the opportunity of billions. By the same token, it also has the potential for severe destruction of rights, such as by misleading users towards catastrophic existential consequences, propagating discrimination, or denying opportunity to those it's deployed on.

Regulating High Impact AI

Approach

For each group affected by AI deployment, the level of impact hinges on the recourse offered to them and how their rights are ultimately affected. If an AI system uses someone's work without their consent, will they have the ability to be notified and remove their work from the system? To be credited for it, or to receive compensation? Or if an AI system automatically denies opportunity to someone who it's deployed on, such as by denying a loan for a home or refusing to cover critical medical expenses, will they have the ability to appeal the decision? And at what cost, time, money, and effort?

Focusing on individual rights helps us to understand how to responsibly regulate AI impact. A **right to existence** includes what has been called "long-term harm" or "existential risk" – annihilation of humanity directly from technology, such as by global extermination through automatically deployed nuclear weapons – but extends to all scenarios where people are killed, now and in the future, directly by AI or indirectly via people who are influenced by AI output, such as medical professionals who subject patients to incorrect medication regimes or inappropriate medical procedures due to AI-generated misdiagnoses. A **right to freedom** and a **right to equal opportunity** roughly correspond to what has been categorized as "current harms" or "short-term risk". Loss of a *right to freedom* includes all situations where AI leads to a person's incarceration, constrained activity, disproportionately larger security requirements, and surveillance. Loss of a *right to equal opportunity* includes situations where AI system performance is disproportionately worse for some subpopulations (a type of "AI unfairness") and where an AI system leads to a loss of compensation or employment.

These rights can be prioritized, and high-risk negative impacts minimized, by the government requiring rigorous documentation of how rights-oriented goals are met throughout development and deployment, following specific protocols. Many of these protocols have already been proposed and have been experimentally implemented in multiple technology companies, such as Model Cards (Mitchell et al., 2019) for machine learning models, and Data Statements (Bender and Friedman, 2018) or Datasheets (Gebru et al., 2018) for AI training and testing datasets. In an AI developmental framework where rigorous documentation is prioritized, an AI system can only continue to be developed and deployed while specific aspects of the documentation are still true. I've worked on this for years and operationalized these documentation frameworks in multiple tech companies, so can explain how they work and why they help.

How it Works: Deep Dive on Model Cards

AI development can be conceptualized at a high-level as following a 4-stage pipeline. Each stage, for each model, in each AI system, should give rise to documentation, creating a paper trail for internal decision-making, government regulation, and auditing.

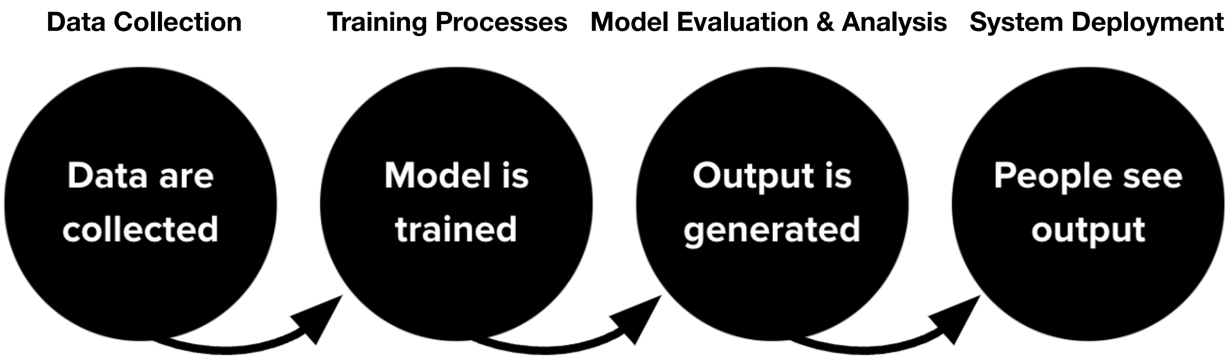


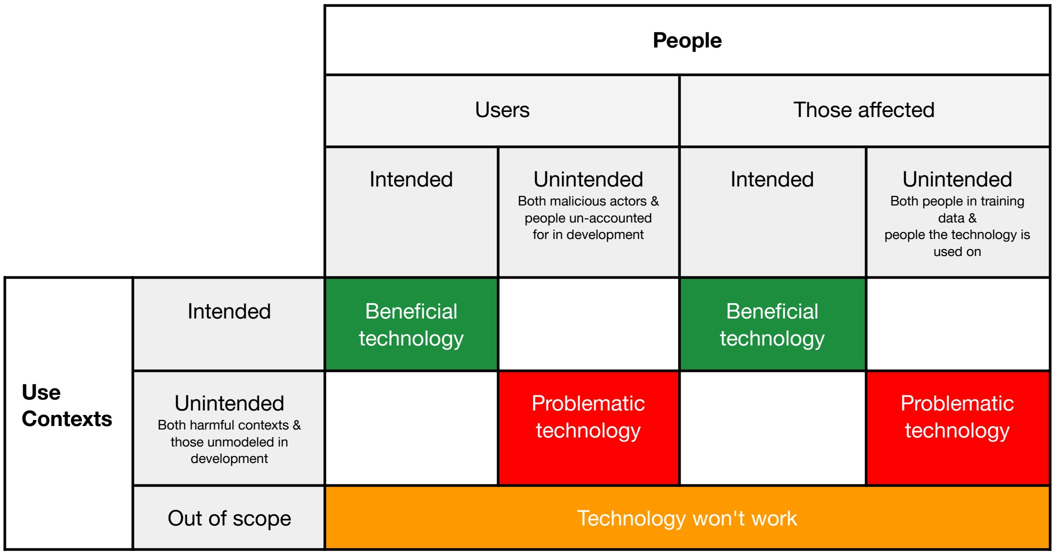
Figure 1. High-level AI model development pipeline.
Each stage can give rise to documentation that demonstrates
meeting different rights-preserving goals defined by legislation.

For the purposes of this meeting, I will primarily focus on the documentation required at the third stage of development, **"Model Evaluation & Analysis"**. I am uniquely qualified for this stage in particular: It is where the documentation framework of "Model Cards" applies, a framework I developed with colleagues that has become a key transparency artifact across the tech industry and that has been referred to in legislation around the world.

Here is what must be documented, presented as ordered steps that can be implemented within a technology organization to best marry technological processes and regulatory requirements:

Step 1. Define the relevant people (stakeholder groups and subpopulations).
Step 2. Identify how each group may be affected in different contexts.
Step 3. Determine the metrics and measurements to evaluate and track the effects on each group (Step 1) in each context (Step 2).
Step 4. Incrementally add training data and evaluate model performance with respect to the groups and contexts (Step 3), measuring and documenting how different inputs affect the outputs.

For Steps 1 and 2, the approach requires crossing **people** by **contexts**. "People" are split into *users* and *those affected*, *intended* and *unintended*. "Contexts" are similarly split into *intended* and *unintended*, as well as *out of scope*. This can be seen as primarily a 2x4 grid, where each cell must be filled out.



<table>
<tr><td colspan="2" rowspan="3"></td><td colspan="4"><b>People</b></td></tr>
<tr><td colspan="2">Users</td><td colspan="2">Those affected</td></tr>
<tr><td>Intended</td><td>Unintended
Both malicious actors & people un-accounted for in development</td><td>Intended</td><td>Unintended
Both people in training data & people the technology is used on</td></tr>
<tr><td rowspan="3"><b>Use Contexts</b></td><td>Intended</td><td>Beneficial technology</td><td></td><td>Beneficial technology</td><td></td></tr>
<tr><td>Unintended
Both harmful contexts & those unmodeled in development</td><td></td><td>Problematic technology</td><td></td><td>Problematic technology</td></tr>
<tr><td>Out of scope</td><td colspan="4">Technology won't work</td></tr>
</table>

Table 1. Foresight in AI chart: Guidance on how to categorize and identify potential impacts. "Unintended" contexts are those the systems have not been developed for: Results are unpredictable. "Out of scope" contexts means those where the system won’t work.

The approach I see to regulating high impact AI empowers developers and auditors to fill out this chart. This means answering the corresponding questions of *what are the use contexts*, and *who is involved in these contexts? What are the intended or beneficial uses of the technology in these contexts? What are the unintended or negative ones?*

Clearly defined subpopulations and use contexts can inform the selection of metrics to evaluate the system, which measure progress in development and assess the system's impacts. Appropriate selection of metrics is critical for guiding the responsible evolution of AI: You can't manage what you can't measure. Evaluation proceeds by disaggregating system performance according to the selected metrics across the defined subpopulations and use contexts. For example, if the goal is minimizing the impact of incorrect cancer detection for women, then **disaggregating evaluation** results by gender is preferable to using aggregate results that obscure differential performance across genders, and comparing results across genders using the *recall* metric (correctly detecting cancer when it is there) may be preferable to a focus on the *precision* metric (not accidentally detecting cancer that isn't there). Such decisions become clear when following the steps described above to identify the key variables at play for an AI system.

There is much more to say about what can be documented at each stage of the development process and the trade-offs and tensions in different metrics and evaluation approaches; this short introduction provides a foundation for model evaluation and analysis that is designed to robustly identify potential impacts of technology.

Why Documentation Helps

The requirement of rigorous documentation throughout the development process and during deployment incentivizes and enforces responsible practices. If you have to list the different contexts of use, you have to think through how the technology is likely to be used, and will further develop with this knowledge in-place. Goal-focused regulation can further help by requiring that specific rights are ensured, such as fairness: If you have to demonstrate fairness across different subpopulations in order to deploy a system, then you will continue to develop the system – and invent ways of ensuring fairness – in order to meet that goal.

Concluding Thoughts

By requiring companies that create high-impact AI to **provide documentation of meeting goals that protect peoples' rights** throughout the development and deployment process, we can connect developer processes and technological innovation to governmental regulation in a way that best leverages the expertise of tech developers and legislators alike, supporting the advancement of AI aligned with human values. This approach is a mix of **top-down** and **bottom-up** regulation for high impact AI: Regulation defines the rights-focused goals that must be demonstrated, such as types of safety, security, and non-discrimination; and the organizations developing the technology determine how to meet these goals, documenting their decisions.

If organizations do not appropriately meet the specified goals or misrepresent their technology, heavy fines that actually affect the bottom line, or disallowing companies to operate in the U.S. market, must be used. This is also an area where additional **whistleblower resources and protections** should be considered, such as the creation of an agency where potential whistleblowers can seek guidance when working at a company that is knowingly creating technology that exposes private information or otherwise harms people outside of military uses.

We do not have to be reactive to the results of technology. By centering on peoples' rights – an approach that government is uniquely qualified for – and focusing on methods for rigorous foresight of the outcomes of technology, we can be proactive, shaping AI to maximally benefit humanity while minimizing harms to our lives, our freedom, and our opportunities.