Written Testimony for
the U.S. Senate Subcommittee on Privacy, Technology, and the Law
Oversight of AI: Insiders' Perspectives
09/17/2024
Dr. Margaret Mitchell

Introduction

Chairman Blumenthal, Ranking Member Hawley, and members of the Judiciary Committee, thank you for the opportunity to testify here today. My name is Margaret Mitchell, and I am here in my capacity as an artificial intelligence researcher and computer scientist who has worked in the tech industry for over 10 years. Within that time, my career has morphed from a primary focus on machine learning and natural language generation to incorporate considerations about the harms and risks of AI technology on people and society. My research today focuses on "ethical AI", where I work on difficult issues at the intersection of AI and society, in order to incentivize responsible development that addresses goals such as fairness, inclusion, and safety.

About ten years ago, while I was a researcher at Microsoft, there was a fundamental shift in technology that brought about what we now refer to as "Artificial Intelligence". At its core, this change was due to neural networks, which had until then been largely theoretical. But by 2014, significant increases in computing power and GPU technology made it possible to demonstrate the real potential of neural networks for many of the tasks we were working on. This included tasks in **natural language processing** research, such as extracting information from text or generating summaries of news articles; it included tasks in **computer vision** research, such as person recognition and object detection; and I was privileged to play a part in, for the first time, connecting computer vision to natural language – so that automatic descriptions of scenes could be generated directly from a smart phone's camera, aiding people who are blind. The possibilities to help people seemed endless.

The catch was that, in contrast to previous approaches, neural network-based approaches require a lot of data and a lot of computing power to work well. And so like many of my colleagues, I tried to amass as much compute as possible to push the limits on what this technology could do – at one point I even had a secret cluster of 8 GPUs in a nearby closet –

and like many of my colleagues, I turned to the newly burgeoning Web 2.0 where content like blog posts, social media conversations, and online photo albums could be fed as data to neural network systems.

However, by 2015, my excitement began to be colored by a deep concern. After I provided my AI system with a series of images depicting a massive explosion, taken from a photo album on the online platform Flickr, the system responded by commenting on how great it was – an effect of the positive sentiment bias in its training data. The images depicted a high-strength fuel-air explosion caused by leaking petrol at an oil storage terminal, which wounded dozens of people.[1] It was in this moment in 2015 that I saw the first glimmers of the future of Artificial Intelligence: Within AI development culture, the connection between what we were developing, which could never truly understand human life, and the potential harms and risks to people once deployed, was being overlooked.

In 2016, I started a new chapter in my career. I had the privilege of joining Google to begin developing AI hand-in-hand with considerations of harms and risk. Because I had seen the criticality of what the training data represents in everything a system may do, I began a deep dive to understand **data biases**. This made clear how significant different data skews are, and how amassing *more* data tended to replicate the same biases and skews, rather than helping to provide a fuller picture. Eventually, with the kindness and support of Google executives, I was able to start a team, which I called "Ethical AI", focused on bringing together interdisciplinary Googlers passionate about social good; and was able to hire a co-lead, Timnit Gebru, who helped to motivate why Google should include the insight of people more aware of the effects of technology on individuals and society.

Throughout my time at Google, my goal was to proactively create beneficial technology that minimized foreseeable negative outcomes. Yet this basic idea of **foresight** – of thinking through all the different ways a technology might evolve, and using this to develop technology that proactively addressed foreseeable risks – was difficult to incorporate into standard development practices. Internal incentives around promotions and raises pushed workers to "launch" (deploy technology and release other artifacts) as much as possible, without collaborating across complementary approaches, incorporating different perspectives, or developing systems that were informed by the impacts of the technology.

Because of this, I realized that some of the responsible practices I had been advocating for, and trying to motivate, might be more successful if I could turn those practices into launches. And so with my colleagues, we introduced "model cards"[2] – artifacts that could be launched that were themselves instantiations of ethical AI practices. Things like thinking through the

[1] Later called "The Buncefield fire".
[2] https://dl.acm.org/doi/10.1145/3287560.3287596

intended use of a system, or comparing the performance of a system on different sensitive subpopulations, such as across race, gender, or age.

Part of my work on rigorous documentation and the role of critical thinking and foresight in development made it clear to me that we must have some amount of **due diligence** before we develop a system, such as research to inform predictions on what the system will be like by reviewing past literature and work from related fields.[3]

I have since continued my work, trying to understand gaps and blind spots in AI development within the tech industry, and develop methods to address them. I have taken on new challenges in understanding when systems might be more "open" or more "closed", and joined AI startup Hugging Face as a researcher and their Chief Ethics Scientist, where, for the first time, I was permitted to study how the AI world I was most familiar with connected to AI policy.

My journey has made clear ways that I believe the government may be able to help shape AI development within the tech sector to be a net positive for the public.

Executive Summary

In the following sections, I explain clear "gaps" I see in responsible AI development that the government might be able to help with, detail possible solutions, and collate specific policy ideas. Additionally, I have provided some information to clarify some common misunderstandings on AI, available in the Appendix.

Gaps

At a high level, the key gaps in AI development practices within the tech industry and AI research that I believe must be addressed for AI to be more beneficial include:

1. Research on **the relationship between model inputs and model outputs**
2. Rigorous **quantification** of data, models, and systems
3. Implementing **due diligence and foresight**, including before development
4. Operationalizing **transparency** in the AI lifecycle, including after deployment

Significant advancements on (1) and (2), and implementation of (3) have not organically emerged from the tech industry or academia; I would therefore like to ask the government's help in catalyzing and incentivizing critical work in these areas. Part of why I am invited today is because I am a leading expert on (4) and can help further explain what might be required from governments to bridge the gulf between industry-internal practices and public accountability.

[3] This fundamental idea underlied my work in the "Stochastic Parrots" paper.

Potential Solutions

These primarily take two forms:

1. Providing research grants and supporting agencies to: (1) scientifically measure how the inputs to a machine learning model affects its behavior (its outputs); (2) create socially informed evaluation protocols; (3) handle private information and prevent harms associated with non-consensual intimate content; (4) create datasets that have appropriate consent, licensing, and compensation where applicable.
2. Requiring documentation and disclosure.

Gaps in the Technology Industry's Approach to Developing and Deploying AI

1. The Relationship between Model Inputs and Model Outputs

There is currently no standard practice of quantifying different aspects of the training data used to train a machine learning model, and no well-developed, mature science analyzing how **inputs** affect **outputs** from the model once trained. This is sorely needed in order to make informed predictions about how AI models may behave as we create new training datasets and train models on them.

It does not have to be a surprise that a model can harm individual rights, such as by generating images of children, exposing private content, or propagating discrimination. We can be informed on whether such content might be generated beforehand: For example, if there are images of children in a model's training data, then there's a good chance that a model will be able to generate images of children. Similarly, it doesn't have to be a surprise that a model can create any kind of world-wide existential harm, such as by activating weapons while generating deceptive explanations that it has not. These scenarios are most likely when there is similar content in the training data.

As such, although there is a divide on whether current harms or long-term existential risk are more important to prioritize, I believe **the solutions are largely the same**. With rigorous tests probing what kinds of *data* results in what kinds of behavior, in controlled conditions, and applied incrementally as the size of a model is increased, we can make well-informed predictions about model behavior now and in the future, and set constraints on the kinds of data that is most appropriate for different scenarios.

The idea is not wholly new: Ablation studies, for example, carefully alter inputs to a system to measure their effects. Yet in the era of "Large Language Models", where data quantity is often considered more important than data quality, these scientific mechanisms are all but forgotten.

Minor Note: There is disagreement on whether a model might generate something that does *not* have corresponding similar contents in its training data – this falls into current discussion on "emergent" properties – but there **is** agreement that if something *is* in the data, a model will be able to do learn it and do something similar: This is the whole point of training a model on data. As such, if the training data contains things like pictures of children, videos of non-consensual intimate content, instructions on how to hack into computer systems, details on constructing bombs, or discriminatory viewpoints and bias, we can make reasonable predictions that a model will produce the same.

It is critical to develop a serious, rigorous science of training data analysis and begin, for the first time in the era of Deep Learning, training models based on data that is curated to represent precisely the intended use contexts of the model.

Policy Suggestion

Because this rather straightforward approach has not developed organically within the tech sector, and has been surprisingly overlooked, I advocate for:

- **Research grants** within academia to catalyze work in this critical area.
- **Requirements** within the tech industry for transparent documentation of data, and especially training data, so that detailed specifics on what a dataset represents can be shared with appropriate parties. In cases of content that cannot be made public, such as private information, intellectual property, and sensitive data, these may be disclosed to appropriate independent parties under NDA.[4]

2. Rigorous quantification of data, models, and systems

This complements Point 1.

Data curation and measurement

Within AI research, the model is king. Priority is given to working on training and inference, and very little attention is paid to the data used to train the models.[5] (Further details on this cultural norm within the tech industry is provided in the introduction.) The goal in collecting training data is amassing sheer volume, not curating content that might be helpful for distinct tasks or that respects peoples' rights and wishes. Curating data by defining **specifically** what we would like a model to learn is an alternative approach to the common approach today, and I believe it

[4] For my most up to date work on how to document a model, see https://huggingface.co/docs/hub/en/model-card-annotated

[5] See for example "'Everyone wants to do the model work, not the data work': Data Cascades in High-Stakes AI", https://dl.acm.org/doi/10.1145/3411764.3445518

will not only result in more respect for the wishes of the general public, but that it will also provide us with key insights on what a model can learn.

As such, just as we *evaluate* AI models, so too can we *measure* AI data. When paired together, we will be empowered with the ability to make informed predictions about how data affects system behavior – and so to develop systems in ways that best align with our goals today and in the future.

Model and System Evaluation

Clearly defined subpopulations and use contexts can inform the selection of metrics to evaluate a model (or system). To ascertain whether a model is "fair", 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 a focus on disaggregating evaluation results by gender 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). Also see my clarification on the "accuracy" metric below.

Policy Suggestion

In addition to those in Point 1, continue to support NIST and the experts therein to facilitate the development of the science of data measurement and standards for data curation and model/system evaluation.

3. Implementing Due Diligence and Foresight

The technology industry largely operates with a cultural norm of reactive approaches and "rapid hindsight". Foresight is generally not rewarded – you're not promoted for the bad headlines that never exist – and this has serious consequences for what technology is developed, whether it's released, and how it's deployed.

Foresight

Some people are better at foresight than others; overlooking this fuels public statements that we "cannot predict" how AI will behave, and thus must deploy it on the public in order to see.

My career has involved sharpening my skills at foresight and helping to teach others. To aid in this, I have developed the following chart. Essentially, foresight within tech development can be informed by 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.

		People			
		Users		Those affected	
		Intended	Unintended Both malicious actors & people un-accounted for in development	Intended	Unintended Both people in training data & people the technology is used on
Use Contexts	Intended	Beneficial technology		Beneficial technology	
	Unintended Both harmful contexts & those unmodeled in development		Problematic technology		Problematic technology
	Out of scope	System won't work			

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 recommend when 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?* As an example, this chart applied to a specific system is available in my Time OpEd on Google's Gemini.

Clearly defined subpopulations and use contexts then can inform the selection of metrics to evaluate the system, as discussed above in Model and System Evaluation.

Due Diligence

ML development and tech culture broadly is relatively *laissez-faire*: Build something first, think about the larger picture afterwards. Incentivizing work with due diligence on social impact, foreseeable harms and risks, and what the public actually needs from AI technology, can fundamentally alter this cultural norm to better align with policymakers' goals.

By "due diligence", I mean the work that a person or organization undertakes to inform what should be built to assist people, and to avoid harm to people or to peoples' ecosystems. It can be used to determine whether a technology should be released or developed at all, and what safety mechanisms to put in place.

Due diligence work includes **background research** on technology – such as reviewing previous work, and work in related fields in order to make informed predictions about how systems may behave before they're released. It includes harms and risks analyses, to determine possible outcomes and identify vulnerabilities before they're exploited. Due diligence should be used to examine the relationship between data inputs and system outputs, as discussed above. When due diligence is in place, inappropriate claims about system capabilities would first be verified, before they're published in press releases that mislead the public[6].

Due diligence is possible with requirements for rigorous documentation. It can be incentivized by **requiring audits** and providing avenues for workers to **responsibly whistleblow**.

Policy Suggestions

- **Require documentation** demonstrating due diligence, including research on the foreseeable harms and risks of a technology under consideration. (This solution also is suggested for (1) above and (4) below.)
- **Support whistleblowing** by creating stronger protections and more available resources to help workers responsibly disclose the development of high risk systems without appropriate transparency.

4. Transparency

Why is transparency important?

Transparency is crucial for addressing the ways in which AI systems impact people. This is because transparency is a foundational, extrinsic value—a means for other values to be realized. Applied to AI development, transparency can enhance **accountability** by making it clear who is **responsible** for which kinds of system behavior; and stops bad practices before they start when the safety of anonymity is removed. This can lessen the amount of time it takes to stop harms from proliferating once they are identified, and **provides legal recourse** when people are harmed. Transparency **guides developers** towards non-discrimination in deployed systems, as it encourages and incentivizes testing for disparate performance and researching best approaches until **fair outcomes** can be reported. Transparency enables **reproducibility**, as details provided can then be followed by others, and the validity of statements can be checked. This in turn incentivizes **integrity and scientific rigor** in claims made by AI developers and deployers and improves the **reliability** of systems. And transparency around

[6] In an official blog post, Google claimed "PaLM has never seen parallel sentences between Bengali and English" and thus it was impressive it could translate between the two languages. As someone familiar with how there is generally not a rigorous understanding of data used in machine learning training, and how machine translation works, I am skeptical that this is true.

how an AI system works can foster appropriate levels of **trust** from users, enhancing **human agency**.

Yet "transparency" need not refer to fully disclosing all information to all individuals; there can be varying levels of transparency, including disclosure to accredited organizations or agencies.

What transparency should there be?

Transparency can be realized via **documentation** and **disclosure.** For documentation, this includes artifacts corresponding to the 4 major phases of AI development:

1. Data preparation
2. Model training & additional system coding
3. Evaluation & Analysis
4. Deployment

Applying this to the typical pipeline of development used within the tech industry brings us a process for transparent documentation that would look something like this:

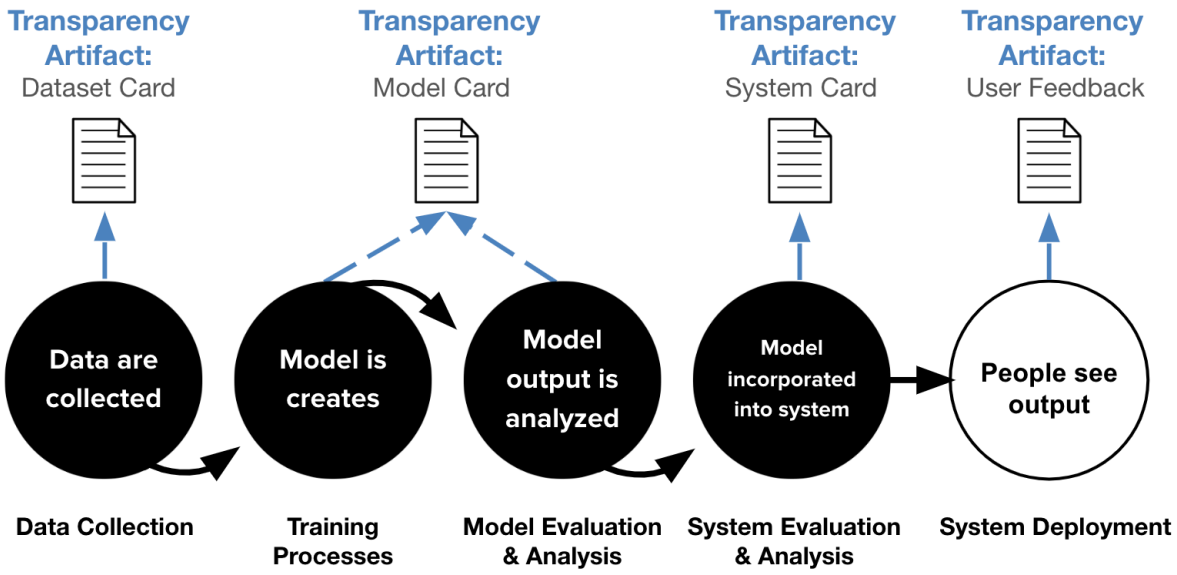

Figure 1. The AI development chain and corresponding transparent documentation

For datasets, datasheets, data statements or dataset cards (which I work on) should ideally be able to assert that the dataset does not contain unwanted or regulated material connected to **foreseeable risk**, and is composed of appropriately consented and licensed data, in addition to further details about its composition that can provide insight into potential model behavior.

For models and the larger systems they're in, model cards and system cards that are understandable to general audiences and provide enough detail for potential users to

understand exactly the situations where the model is most appropriate. This includes specifying intended use, bias, risks, and limitations; disaggregated evaluation across different contexts and subpopulations; and social impact assessments informed by practices such as red-teaming. Please see this resource for my most up-to-date work on how to create model cards.

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 machine learning (statistical) 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.

Policy Suggestion

For large deployed systems, require documentation demonstrating due diligence on harms and risk analyses before development. For companies that are in scope, require documentation that can be audited detailing the development process. This same type of documentation is suggested to address (1) and (3) above.

Summary of specific Policy-Relevant Ideas

This collates policy-relevant ideas I've shared here, in publications, and in D.C. congressional events, including at the Senate Insight Forum on High Impact AI and listening sessions with the House of Representatives[7]. These come with the strong caveat that I am not a regulator and do not have as full a picture of the regulatory landscape as those here today. My goal is to try to offer some ideas that might be helpful to inform real legislation.

Policy Scope

I have found it useful to navigate ethical AI by centering 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. It includes the following (non-exhaustive and not mutually exclusive) stakeholder groups:

[7] Including the Democratic Caucus AI Policy Panel, "AI and the Future": https://m-mitchell.com/statements/StatementforDemCaucus.pdf and the Congressional AI Academic Roundtable: https://m-mitchell.com/statements/RoKhannaRoundTable.pdf

- **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 tech 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)

By defining relevant subpopulations and stakeholder groups in AI, it is possible to derive the benefits, harms, and risks to different people in different contexts, identifying potential for positive and negative impact. For details on each group and how regulation may apply to them, please see my Senate AI Insight Forum statement.

For the stakeholder group of *AI-affected* (all of us), consider policy that concerns **children**, **weapons/violence**, **non-consensual intimate content**, and **systems that can take actions in the real world**.

Providing Funding

I believe it could be beneficial for the government to provide funding to address clear gaps in the AI ecosystem. Because dedicated development on several critical topics have not received much attention within academia and the technology sector, and/or because it has been too expensive for relevant entities, it would be useful for the government to help realize different AI practices that are relevant to protecting Americans.

- **Provide research funding for key issues that are currently "gaps" in AI development.**
    - Funding from NSF, NIH, and DARPA have all fundamentally shaped the state of the art in AI. For example, the internet as a useful resource for knowledge has largely been shaped by NSF projects. DARPA grants for projects on translating between languages have pushed the boundaries of machine translation technology such that we can now have relatively fluent conversations with translation between different languages
    - This might include funding for research on:
        - **Privacy**, such as Identifying, redacting, and pseudonymizing personally identifiable information (PII) and other types of personal data. Also see Clarifying Common Misconceptions - Privacy for further details.

        - **Provenance**, such as watermarking of text and images. While visible watermarking is quite possible, invisible watermarking that is robust to different types of alterations is an open research area.[8]
        - **Environmental efficiency**, such as energy-efficient and water-efficient AI. One of the largest contributors to AI's environmental impact is what happens in data centers where models and data are stored (which then are accessed via Cloud infrastructure). This includes everything from the cement and steel used in the buildings to the cooling systems requiring electricity and water to keep servers from heating up. For a primer, please see: https://huggingface.co/blog/sasha/ai-environment-primer
- **Continue funding the National Institute of Standards and Technology:** NIST plays one of the most critical roles in helping AI to be informed by the real needs of Americans, tasked with the tough challenges of defining standards for AI development, measurement and evaluation. Increased funding should enable and empower NIST to do this as well as possible.
- **Continue funding and supporting the National Artificial Intelligence Research Resource (NAIRR):** The NAIRR facilitates the democratization of responsible AI[9] development by drastically increasing the participation of US organizations in AI research and ensuring that important research also happens outside of the few best-resourced companies.
- **Support social impact assessments**, particularly from experts in relevant fields.
    - This has direct relevance to practices such as red-teaming, where it is critical that people in different high risk domains dealing with medicine, weapons, etc., can stress test systems directly before they are deployed at scale.
    - It would be extremely helpful if the government can help with social impact assessment testing, either through creating an independent group or providing funding for this work.

Requiring testing and documentation

- Entities should demonstrate **due diligence** on foreseeable outcomes of technology before it is deployed.
    - See the "Stochastic Parrots" paper I co-authored for an example of due diligence on harms and risks of language models.

[8] For a list of open source approaches to provenance and other deep fake technology, please see: https://huggingface.co/collections/society-ethics/provenance-watermarking-and-deepfake-detection-65c6792b0831983147bb7578

[9] Further details available at https://huggingface.co/blog/assets/92_us_national_ai_research_resource/Hugging_Face_NAIRR_RFI_2022.pdf

- Entities should **evaluate systems** in a way that demonstrates fair treatment across protected groups.
    - This idea is fundamental to the Model Cards endeavor.[10]
- Instead of requirements on what people and organizations must *do*, it may be helpful to place requirements on what people and organizations must *show*, such as fair performance across protected groups, called "machine learning fairness".
    - This also helps to catalyze (not stifle) innovation; and to marry internal self-regulatory processes to external regulation.
- Note that not all documentation needs to be fully public, such as when it would infringe on intellectual property; disclosure to appropriately accredited parties dependent on the content is warranted.

Requiring disclosure of AI-generated media

Media that can be used to misrepresent or maliciously represent people and events should be identifiable as an unfaithful depiction of the real world. Everything from edited photography to video games may fall under this definition; AI-generated content is one area where misunderstandings and disinformation are a significant risk. Operationalizing transparency for AI-generated content requires:

- It should be disclosed when a user is interacting with AI-generated content.
- Deployers must process the disclosure to make it accessible to people who are exposed to the content.
- Platforms must ensure that content circulating on their platform are appropriately flagged as such to users.

Creating further mechanisms for tech whistleblowing

The tech industry could benefit from stronger protections and more available resources would help workers for responsible disclosure to a government-supported party, such as disclosure of inappropriate development practices or false public information. There are currently limited options for who tech workers can talk to when they see their company engaging in harmful practices, and few resources. Just as the financial sector has the SEC for financial whistleblowing, and Acts have been passed to support whistleblowers in Motor Vehicles, False Claims, Pharmaceuticals, Medical Practice, etc., so too should there be specific support for whistleblowing in the technology sector. (Note, too, that although whistleblowing is protected, it is trivial for companies to retaliate for whistleblowing and legally claim it is not retaliation – especially given the common At-Will employment clause – making whistleblower protections currently virtually meaningless, as far as I can understand.)

[10] Original paper: https://dl.acm.org/doi/10.1145/3287560.3287596
Current recommended approach: https://huggingface.co/docs/hub/en/model-card-annotated

Supporting the development of auditing & compliance protocols

Companies creating AI should transparently demonstrate compliance to standards and law. In some cases, this might be via closed-door auditing (similar to financial auditing), where an external party under non-disclosure agreement (NDA) can verify compliance in cases where content cannot be fully shared publicly. Areas of development that should be transparently compliant include:

- Training data: Adherence to licenses, copyright, and standards on malicious content.
- Testing protocols: Appropriately robust evaluation for foreseeable uses, including misuse, malicious use, and disparate impacts across different subpopulations (such as by race, gender, age, ability status, etc.).
- System analysis: Tracking the relationship between training data inputs and system outputs (discussed above) to construct a reasonably detailed report on the system's scope of behavior.

Strengthening data rights

Creators should not have their work uncompensated and uncredited. Previous law did not provide for AI-relevant recourse for people who share content online, and so we must create new approaches that are sensitive to the fact that creators' wishes are not being respected.

Supporting C2PA, EXIF, and similar metadata standards

To enable creators to embed content with their ownership. For further details on the state of the part, please also see the collection of tools relevant to provenance I have put together with colleagues.

Providing recourse for individuals harmed by AI

When AI harms someone, what should they do, and who is at fault? I hope regulation can help to provide avenues of recourse for people who may be harmed by deployed AI systems.

Other Relevant Statements, Articles, and Publications

Statements

Democratic Caucus AI Policy Panel, "AI and the Future", discussion with leading experts on the risks of Artificial Intelligence (AI), July 26, 2023:
https://m-mitchell.com/statements/StatementforDemCaucus.pdf

Senate Insight Forum, "High Impact AI", November 1, 2023:
https://m-mitchell.com/statements/SenateInsightForum-HighImpactAI.pdf

Congressional AI Academic Roundtable, February 15, 2024:
https://m-mitchell.com/statements/RoKhannaRoundTable.pdf

Presentation to Congressional AI Caucus, February 14 & 15 2024:
AI Caucus Copy of An Ethics-Informed Approach to AI

Articles

"The Pillars of a Rights-Based Approach to AI Development", Tech Policy Press, DEC 5, 2023
https://www.techpolicy.press/the-pillars-of-a-rightsbased-approach-to-ai-development/

"Ethical AI Isn't to Blame for Google's Gemini Debacle", TIME, FEB 29, 2024
https://time.com/6836153/ethical-ai-google-gemini-debacle/

"The CLeAR Documentation Framework for AI Transparency: Recommendations for Practitioners & Context for Policymakers"
https://shorensteincenter.org/clear-documentation-framework-ai-transparency-recommendations-practitioners-context-policymakers, May 21 2024

"How to Stop Deepfake Porn Using AI", Teen Vogue, July 12, 2024
https://www.teenvogue.com/story/how-to-stop-deepfake-porn-using-ai

The Environmental Impact of AI – Primer, September 2024
https://huggingface.co/blog/sasha/ai-environment-primer

Human-Centric Computer Vision Landscape for Single Images, Medium: A resource for understanding the different levels of "human centric" technology
https://medium.com/@margarmitchell/facial-analysis-or-recognition-c7c5554a43dc

Relevant Research Publications

Model Cards for Model Reporting

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

Measuring Data

Appendix: Clarifying Common Misunderstandings

Below are positive assertions about what I understand to be true that contrast with current misunderstandings.

"Accuracy" is not always how models and systems should be evaluated.

"Accuracy" is a term that colloquially refers to evaluation score, but within machine learning, it is one specific type of metric – there are many others. Different metrics are more useful and informative than others for different systems and contexts. For example, for cancer detection, *finding cancer when it is there* may be more important to prioritize than *not finding cancer when it is not there*. The first is measured with "sensitivity" (also called "recall"); the latter is measured with "precision" and "specificity". Please do not require the "accuracy" metric specifically within legislation, but instead the metrics that are most relevant for the tasks the system is used for.

"Data" in AI is an umbrella term for many different things.

When someone is transparent about "data" that doesn't mean they are being transparent about *all* data involved in a system's design. Similarly, when someone says a solution requires no data – such as RLHF or zero-shot learning – this leaves out the fact that these solutions require operating on models that have already been pre-trained on a lot of other data.

To clarify, there are several types of data used in machine learning, including:

- Training data (what the model is trained on)
- Validation data (what the model is tested on during training)
- Testing data (what the model is tested on once training is complete)

Increasingly, in the case of chat systems, there is also what might be called "system interaction data" – there is not currently a consensus term for this – which is the content produced when a deployed system interacts with a user (such as in ChatGPT conversations).

Training data is currently split into a few types, "pre-training data", and "finetuning data". "Pre-training data" is the bulk of the data; it's the fundamental basis of a machine learning model. "Finetuning data" is the data a model may be further trained on, once it has already been pretrained, in order to work for a specific task. Companies may speak to how they are being transparent about their "training data", when what they are referring to is the finetuning data – not the bulk of the data that forms the basis of the model.[11] Similarly, they may say a model requires "no data" when instead they are referring to no *fine-tuning data*.

"General purpose" is a myth (or a way of not being clear).

Systems are not trained to be general-purpose as such; they are trained on data, and that data represents a multitude of different tasks. Currently, when someones calls a system "general", it simply means that the person saying it is not specifying what the different tasks are, not that the system will work for everything.

This is distinct from more traditional task-specific systems, where training data must correspond to the task that a system is intended to do. The specific tasks that might comprise "general" are not made clear; generality is meant to be produced from the vast amount of data that systems are trained on, where multiple tasks are represented as a side effect of data collection across different sources. As such, the goal in collecting training data is amassing sheer volume, not curating content that might be helpful for distinct tasks. This introduces many issues referenced throughout this statement.

"Transparency" does not mean competitors must know trade secrets.

Transparency can mean everything from (1) sharing details about training data to a trusted set of individuals under NDA to (2) making training data fully available for the entire public. When regulators ask companies for transparency about training data, it's important not to get pulled into a discussion about the problems with (2) when the solution lies in (1).

"Privacy" protection in data does not work well.

Although companies may employ the best-in-class approaches to personally identifiable information (PII) redaction, "best-in-class" does not work well. Those interested in working on this issue could use resources to significantly improve this. One important piece of this puzzle to be aware of is that some types of PII are much more easily detected than others – while some kinds may be trivially removed without issue, others may not. For example, email addresses follow a fairly regular pattern unique to email addresses and not found in other kinds

[11] A twitter thread I wrote on this issue is available here: https://twitter.com/mmitchell_ai/status/1646242689862729728/

of text (such as having an @ symbol), which make them easy to detect and remove. PII such as numbers – such as social security numbers, phone numbers, or (even harder) strings of numbers that might be used to identify someone in a personal record such as a medical record, are easily confusable in text with things like math, page numbers and dates. These kinds of PII have a **high false positive rate**, meaning that removing them can mean removing other critical content that might be desired for a system to be able to perform various tasks.

PII is also an American concept, defined by the department of defense (link), and state-of-the-art privacy removal might successfully identify social security numbers while missing all other kinds of national identification numbers (for example). For a relatively full list of what kinds of private information there is support for, see Microsoft's Presidio.

Machine learning is not just like human learning.

Minimally, humans are made of meat and protein; computers are made of silicon, plastic, etc. Any comparison between human and machine learning is thus an analogy: The two systems are objectively and undeniably not identical. Which analogies to use are a matter of personal preference, and appear to be shaped by misunderstandings (or intentionally creating misunderstandings).

Another word relevant for communicating about learning is "fit": When a model "learns", that means it is *fitting* to the data it sees during training; an analogy would be like pressing dough into a mold.

Figure 2. Example of "fitting" to a mold.
Source: https://thriftdiving.com/vintage-diy-clay-mold-appliques-furniture/

If you don't think that how a human "learns" is by doing something like "fitting" to a mold, then you can see how the "learning" analogy breaks down: The polysemy of the word "learn" is being abused to mischaracterize the nature of machine learning.

The ability to speak human languages does not magically emerge.

There is active debate in the AI research community on what properties of AI systems are "emergent". A combination of the media and corporate comms has misled the public towards believing that, for example, an AI system could magically translate English to Bengali, without ever having seen Bengali. This is simply not so. There is not agreement in the AI community that any behaviors are emergent. We in the AI community are still trying to pin down specifically what that term means and how it might be measured.[12]

[12] See for example Schaeffer et al, 2023, "Are Emergent Abilities of Large Language Models a Mirage?" https://papers.neurips.cc/paper_files/paper/2023/file/adc98a266f45005c403b8311ca7e8bd7-Paper-Conference.pdf