

# **The Data Governance Gaps--Generative AI and Data**

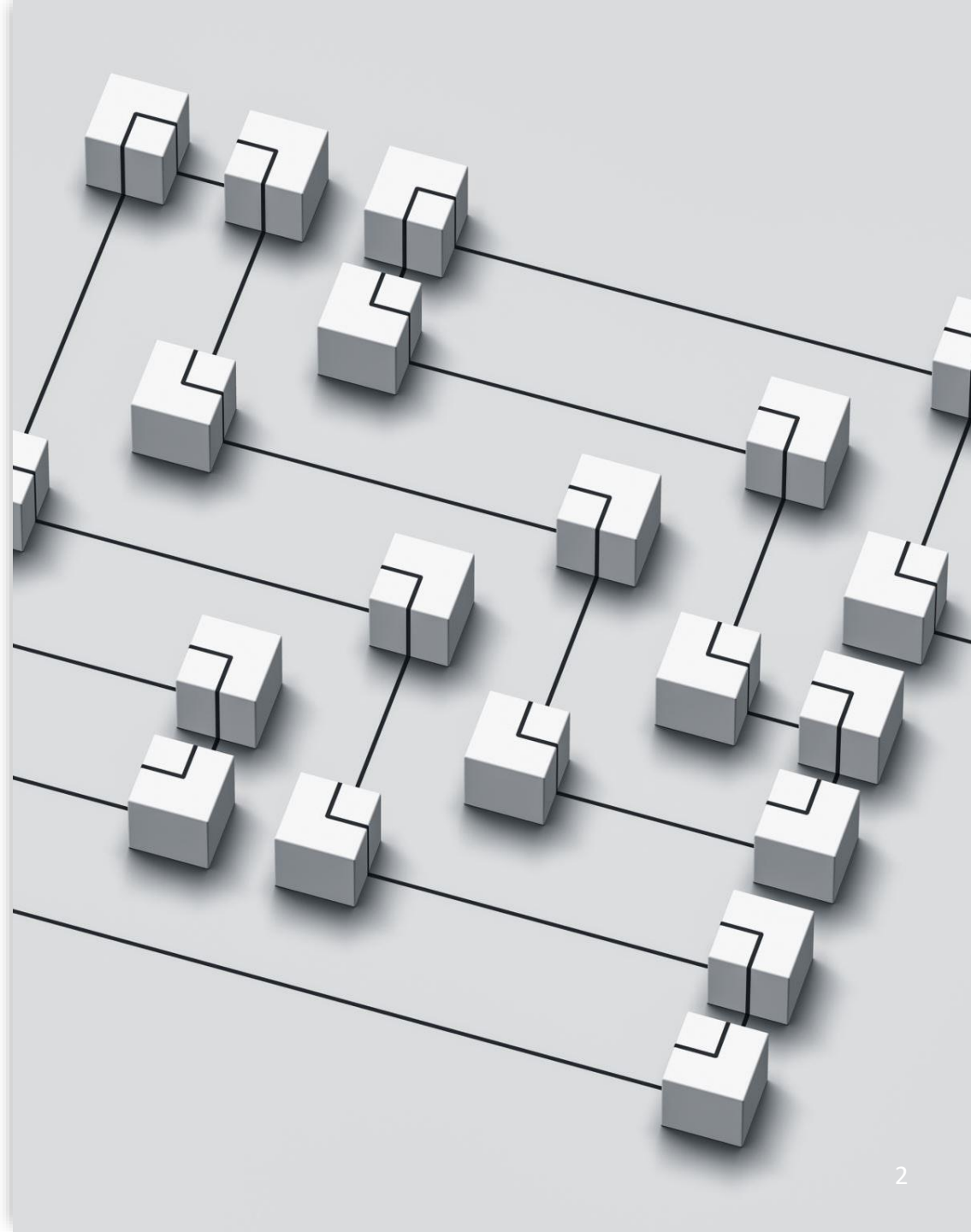
**Susan Ariel Aaronson, Ph.D.**

**Research Professor GWU, Director,  
Digital Trade and Data Governance  
Hub, and co-PI Trustworthy AI  
Institute for Law and Society**

For reuse, contact [saaronso@gwu.edu](mailto:saaronso@gwu.edu)

# NIST-NSF TRAILS Mission

- *TRAILS designs, develops, deploys and models a new, more accountable, inclusive, and participatory approach to AI. We use convenings, training, and interdisciplinary research to empower users to make sense of and participate in the development and governance of AI systems.*





# Presentation thesis

- Data governance is an effective route to governing generative if we can review and address the data governance gaps.
- But policymakers and the public need to think carefully about the international implications of their policy choices.



# Overview: An International Perspective on Generative AI

- AI is everywhere.
- [Generative AI](#) is also [now everywhere](#). [It is built on neural networks which use large language models \(LLMs\) to generate something new.](#)
- Issues: How did the firms creating these LLMs get their data to train? Did they follow internationally accepted rules regarding copyright and personal data protection?
- Is that data sets accurate, complete and representative of the world, its people and their cultures?)
- Growing numbers of policymakers and activists in the developing world are worried about [expropriation](#) of resources (data) and paying additional rents to big tech companies located in the West and China.



# Leads to Two big questions:

What are policymakers trying to govern?

Which of these governance issues requires international coordination?



The **high risks of the technology?**  
(autonomous weapons?)



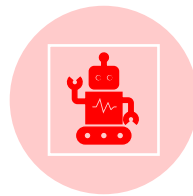
Liability—**who is liable** if the technology hurts a person or persons?



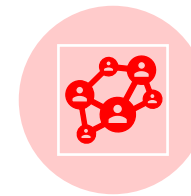
The **uses of the technology** (can it be used to provide therapy or legal advice?)



The **business practices** or policies that firms use to supply the technologies (e.g. free services in return for personal data)?




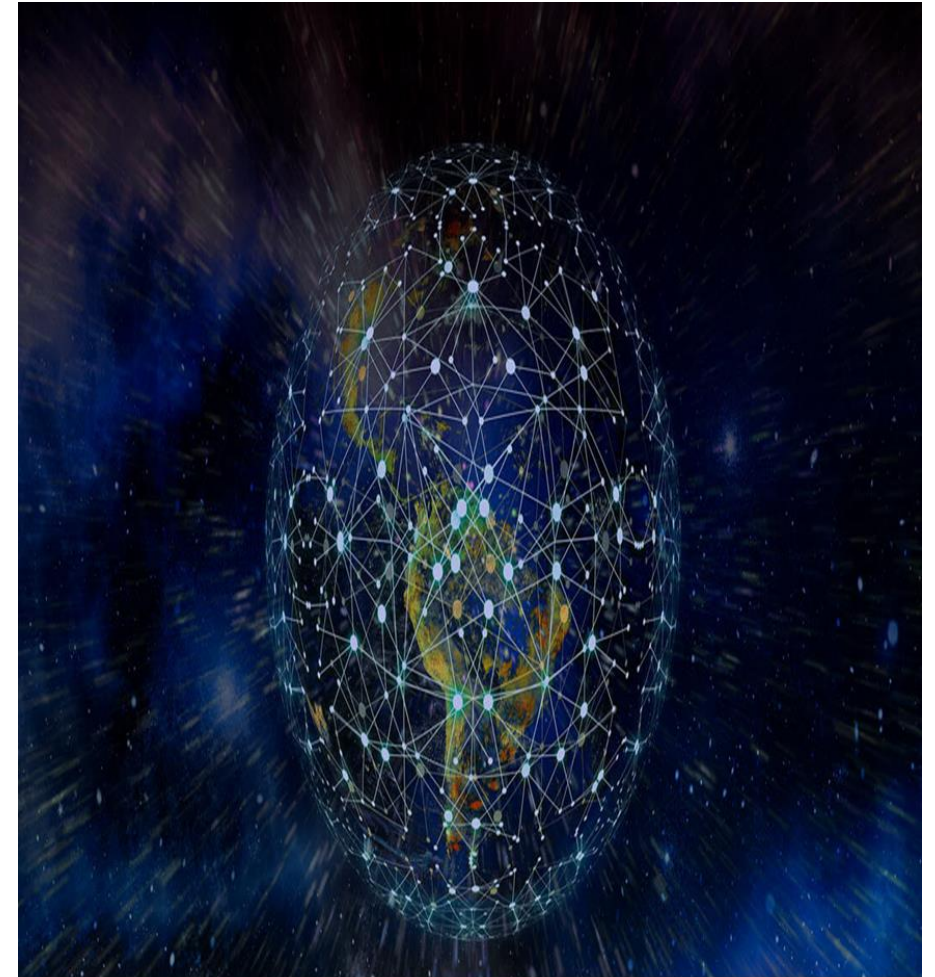
**The data** utilized to create the generative AI service? (e.g. web scraping Linked-In)



Which of these need to be coordinated internationally? How can they be coordinated?

# AI in the world

- Global market for AI huge and growing
- 2 countries--China and US -94% of all AI funding, US 73% of generative AI firms. These platforms not only lead on AI, but control data collection through control of platform services, submarine cables and satellites, data storage and data analysis.
- These firms often use open-source methods, but in general they rely on trade secrets to protect their algorithms and to control and reuse the data they analyze.
- Dominance of these platforms  perception
- AI markets =unfair
- Generative AI could enhance these concerns except that several governments (UAE, France, UK?) have or are establishing open source federally funded LLM chatbots







To understand AI, we need to talk about data. DATA is:

**Easy/cheap to store, move across borders**

- **Easy to share, reuse but increasingly hoarded by governments and firms.**
- **Can be simultaneously a commercial asset and a public good.**
- **There are many types of data**
- **Essential to national security and economic growth**
- **Foundation of wide range of services yet no one knows how to govern the use and reuse of data.**



# What is the data supply chain for Generative AI?

- **Proprietary Data**

data protected by trade secrets or other IPR;

data collected by or purchased by the AI designer/deployer

- **Web scraped data including:**

- **Proprietary, personal, and open-source data where individuals created and provided content.**



# Questions Raised by Data Supply Chain for Generative AI

- 
- How web scraping may affect copyrighted data
- How web scraping may affect individuals and groups who are supposed to be protected under privacy and personal data protection laws.
- How web scraping revealed the lack of protections for content creators on open access web sites; and
- How the debate over open and closed LLMs may affect individuals and firms which hold data and reveals the lack of clear and universal rules to ensure the quality and validity of datasets.

For reuse, contact [saaronso@gwu.edu](mailto:saaronso@gwu.edu)



This Photo by Unknown Author is licensed under [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/)

# Firms Don't say much about Data Provenance

- Meta's LLAMA 2 - <https://arxiv.org/pdf/2307.09288.pdf>  
LLAMA 1 - <https://arxiv.org/abs/2302.13971>

Open Ai's GPT-4 - <https://arxiv.org/pdf/2303.08774.pdf>  
GPT-3 - <https://arxiv.org/abs/2005.14165>  
GPT-2 - [https://d4mucfpksywv.cloudfront.net/better-language-models/language\\_models\\_are\\_unsupervised\\_multitask\\_learners.pdf](https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf)  
GPT - [https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language\\_understanding\\_paper.pdf](https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf)  
BLOOM - <https://openreview.net/pdf?id=UoEw6KigkUn>

◦



# How have governments responded to *generative AI*?

- [Italy banned, then rescinded the ban on Chat GPT](#) after the company made changes.
- [Canada's privacy](#) commissioner is investigating Chat GPT
- [Spain](#) opens an investigation into Open AI
- Britain is levying fines on personal data [misuse by Chat GPT](#)
- [European consumer groups](#) call for an investigation of Chat GPT, while [European data protection](#) authorities agreed to set up a task force to cooperate and exchange information on the relationship between European laws and such chat bots.
- China puts forward new rules for LLMs like Chat GPT(translated by DIGIChina ) as of 1013, including [security requirements](#) for AI training data including ways to avoid IPR violations and to seek consent for personal data
- [FTC investigation](#) of Open AI and the data underpinning Chat GPT
- [France-CNIL rules on use of personal data](#) for chatbots
- G-7 Hiroshima Process working towards shared generative AI principles



# Review- Data Supply chain and data governance

- Enforcement problems:
- Copyright, personal data because data supply chain is opaque.



- **Governance gaps--**
- **Those who create content on websites like Reddit or Wikipedia have no protections for their work**
- **No formal protections regarding reuse of data from government websites of data collected or funded by government. Could this lead to a few firms/chatbots become the source of information on what government is doing and on science?**

# Larger international context and implications



# Benefits to Global Science of Generative AI



- Huge potential to help solve wicked problems.
- Indirect incentive to data sharing, as sharing data would allow more people and more countries to achieve the benefits of AI and open science.
- Could facilitate more access to open data and inform more people about the benefits of open data and open science.
- May prod scientists, developers, designers and deployers to invest time and resources into data provenance and data stewardship –could lead to better data standards and better use of data.
- May require a rethink of how we produce and share scientific data. Is it time for this, given distrust of science and experts?



# Costs to Global Science of Generative AI



- Use of generative AI chatbots could undermine trust in science as hallucinations distort what is real and trustworthy.
- Researchers may be less willing to share data without compensation and without some form of accountability that their work won't be misrepresented or misused.
- Policymakers may hoard all types of data, may be reluctant to share.
- Policymakers may be forced to choose between innovation and other priorities. As example, [China chooses censorship](#) over accurate generative AI based on global datasets.

# Larger debate: US tradition of openness threatened by competition with China

- Some argue openness empowers China as China [appears to embrace open source](#).
- But China is generally closed (Information Law, Great Firewall). Nonetheless, US should not become more like China to compete with China in AI.
- Openness could be a source of comparative advantage.
- Do we really want to limit openness and data sharing?



The background of the slide features a vertical gradient from light blue on the left to a warm orange-brown on the right. On the right side, there is a complex, abstract pattern of thin, light blue lines that form a web-like structure. Scattered throughout this web are numerous circles of varying sizes. Most of these circles are in shades of blue, ranging from light to dark. At the bottom of the image, there is a cluster of circles in shades of green and teal, also of varying sizes, which appear to be part of the same web-like structure.

# Thank you

I welcome your questions and comments