

Berkeley

Center for  
Decentraliz

Verizon Event ...



## Keynote Speaker

### Building an Open, Responsible AI Economy

**Dawn Song**  
Professor  
UC Berkeley



Berkeley  
University of California  
Powered by Zoom



# *Building an Open, Responsible AI Economy*

Dawn Song  
UC Berkeley

Berkeley

Center for Responsible,  
Decentralized Intelligence



Center for  
Human-Compatible  
Artificial  
Intelligence

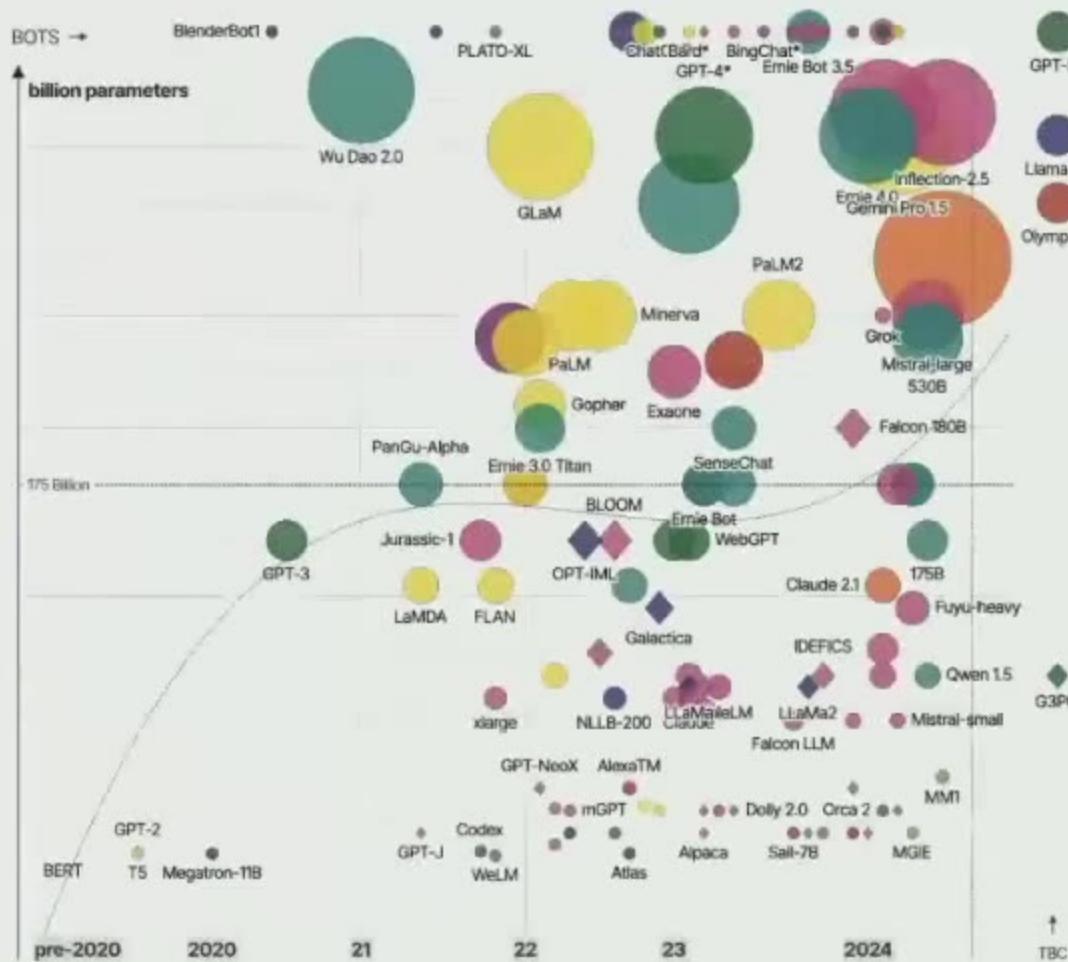


# Exponential Growth in LLMs



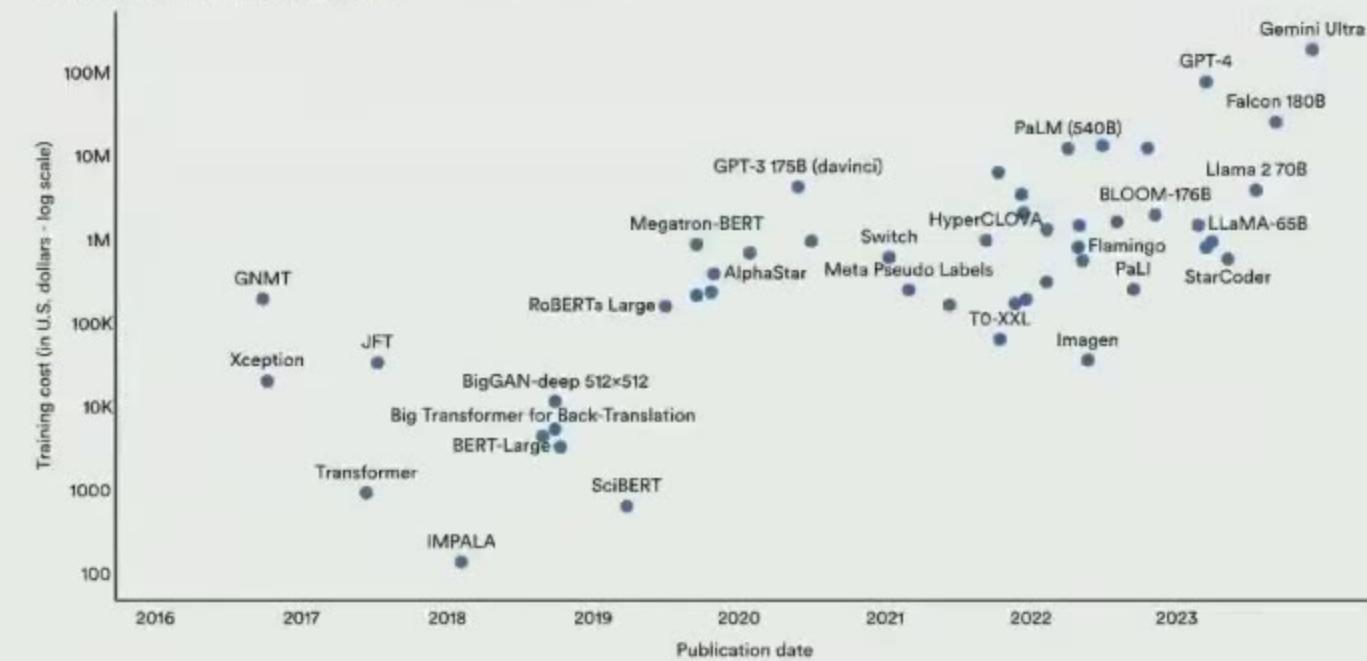
## Large Language Models (LLMs) & their associated bots like ChatGPT

● Amazon-owned ● Anthropic ● Apple ● Chinese ● Google ● Meta / Facebook ● Microsoft ● OpenAI ● Other

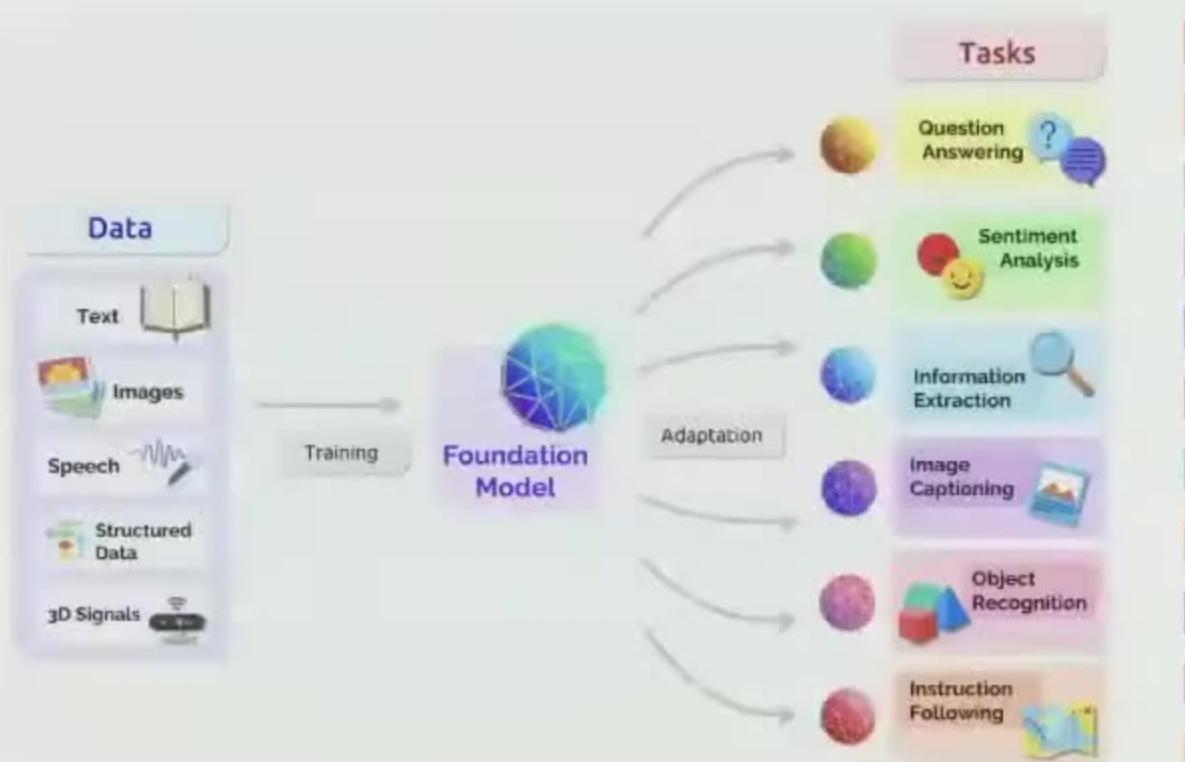


## Estimated training cost of select AI models, 2016–23

Source: Epoch, 2023 | Chart: 2024 AI Index report



# Powering Rich New Capabilities



<https://arxiv.org/pdf/2108.07258.pdf>

Q&A <small>Answer questions from various domains</small>	Grammar correction <small>Correct grammar in 100+ different languages</small>	Broadcastsheet creator <small>Create broadcastsheet from news stories</small>	JavaScript helper chatbot <small>JavaScript helper bot for web communication</small>
Summaries for a 2nd grader <small>Generate different and suitable summaries</small>	Natural language to OpenAI API <small>Convert text to suitable and different AI API calls</small>	ML/AI language model tutor <small>Provide language support for research purposes</small>	Science fiction book list maker <small>Generate a list of science fiction books</small>
Text to command <small>Convert text to command</small>	English to other languages <small>Translate English text into various languages</small>	Tweet classifier <small>Identify tweets from various categories</small>	Airport code extractor <small>Extract airport codes from text</small>
Natural language to Stripe API <small>Convert text to Stripe API's secret key</small>	SQL translate <small>Translate between various SQL engines</small>	SQL request <small>Send raw SQL request</small>	Extract contact information <small>Extract contact information from a resume</small>
Parse unstructured data <small>Extract useful knowledge from unstructured text</small>	Classification <small>Classify text into categories or sub-categories</small>	JavaScript to Python <small>Convert various JavaScript code snippets to Python</small>	Friend chat <small>Engage in real-time communication</small>
Python to natural language <small>Explain a piece of Python code in natural language</small>	Movie to Emoji <small>Convert movie titles into emojis</small>	Mood to color <small>Find the best color palette for mood</small>	Write a Python docstring <small>Autogenerate of basic documentation for Python</small>
Calculate Time Complexity <small>Estimate time complexity of a function</small>	Translate programming languages <small>Translate code from one programming language to another</small>	Analogy maker <small>Discover meaningful analogies</small>	JavaScript one line function <small>Turns functions into one-line equivalents</small>
Advanced tweet classifier <small>Identify different tweet categories</small>	Explain code <small>Explain a piece of code</small>	Micro horror story creator <small>Generate micro horror stories</small>	Third-person converter <small>Converts first person to third person</small>
Keywords <small>Extract keywords from a document</small>	Factual answering <small>Provide factual answers to common questions</small>	Notes to summary <small>Take notes and generate a summary</small>	VR fitness idea generator <small>Find ideas for VR fitness activities</small>
Ad from product description <small>Extract ads from product descriptions</small>	Product name generator <small>Generate product names from descriptions</small>	ESRB rating <small>Get ESRB rating for a game</small>	Essay outline <small>Generate an outline for an essay</small>
TL;DR summarization <small>Summarize long documents in 100 words</small>	Python bug fixer <small>Fix bugs in Python code</small>	Recipe creator (at your own risk) <small>Generate recipes from a list of ingredients</small>	Chat <small>Engage in real-time communication with an AI model</small>

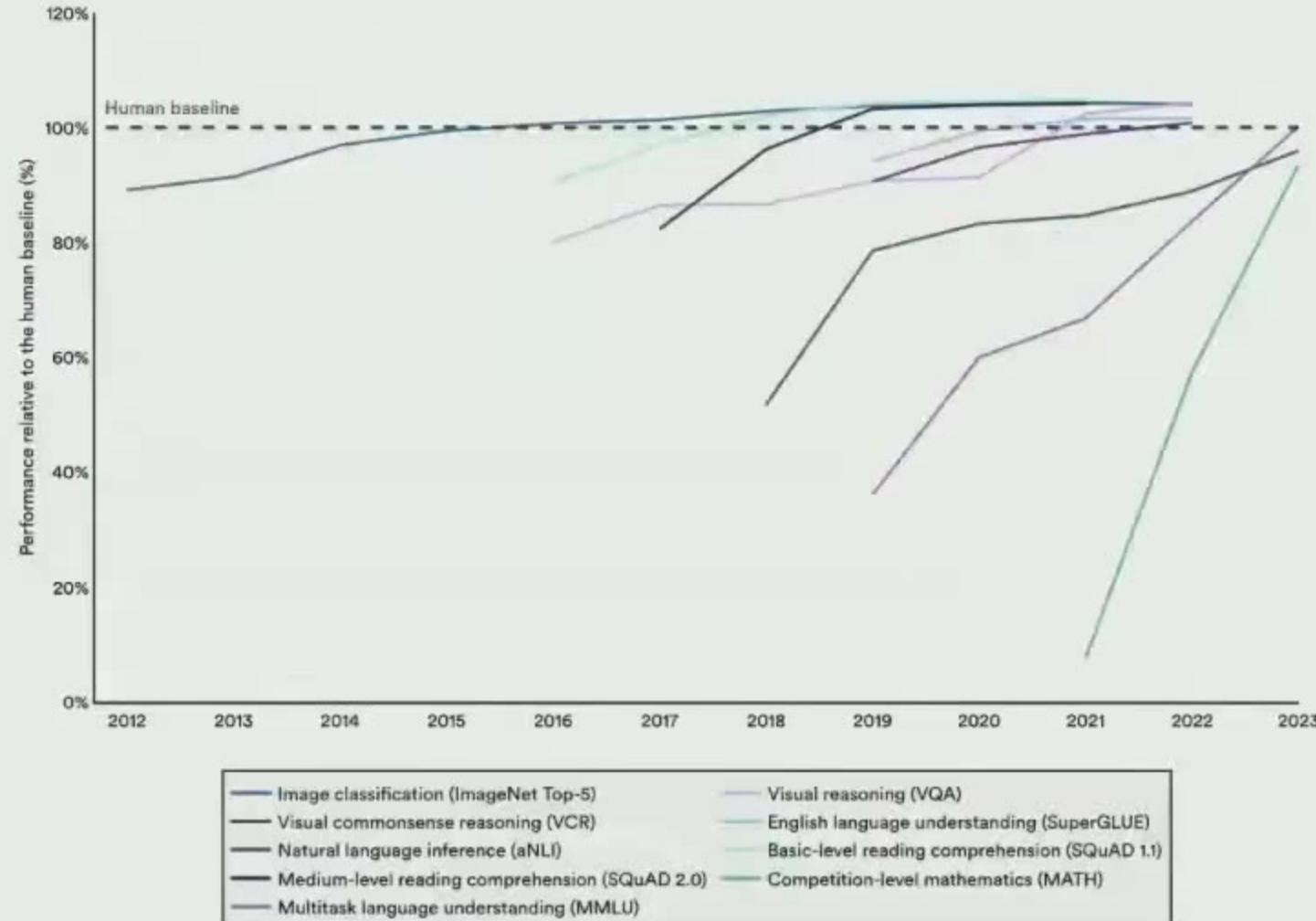
Source: openai

# Rapid Advancement on AI Model Performance

Verizon Event ...

Select AI Index technical performance benchmarks vs. human performance

Source: AI Index, 2024 | Chart: 2024 AI Index report

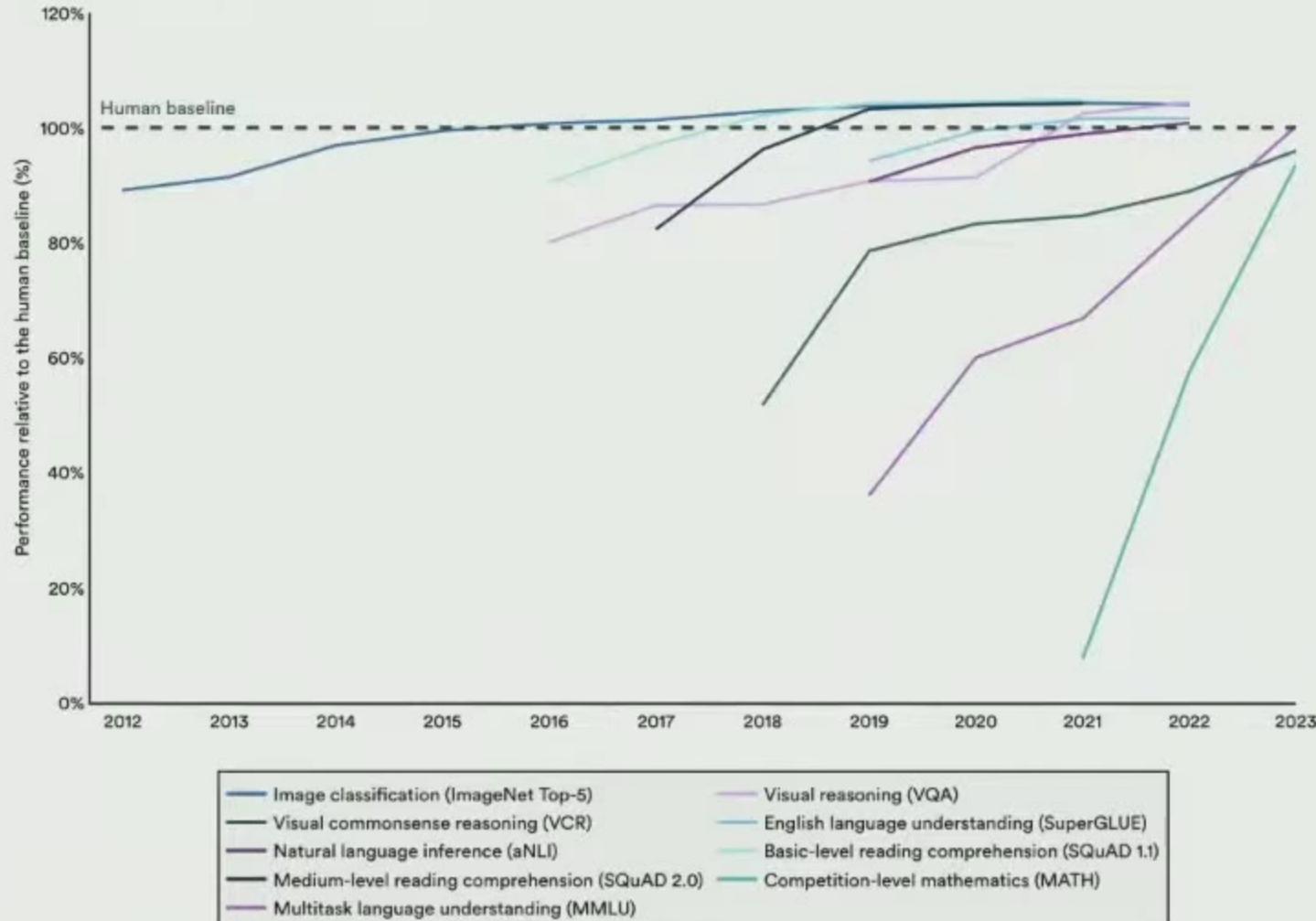


# Rapid Advancement on AI Model Performance

Verizon Event ...

Select AI Index technical performance benchmarks vs. human performance

Source: AI Index, 2024 | Chart: 2024 AI Index report



# The AI Era



- Agents
  - Personal agent/assistant intermediate most communications with external world
  - Diverse agents for solving different tasks
- Decentralized autonomous agents
  - Interact & act in a decentralized manner

# Challenges for the AI Era



- Responsible, safe AI
  - Ensuring trustworthy AI
  - Mitigating misuse of AI
  - Mitigating loss of control
- User-controlled, open architecture & protocol for innovation instead of platform-controlled
- Responsible, fair data use & value attribution instead of platform-centric economy



# Challenges for the AI Era

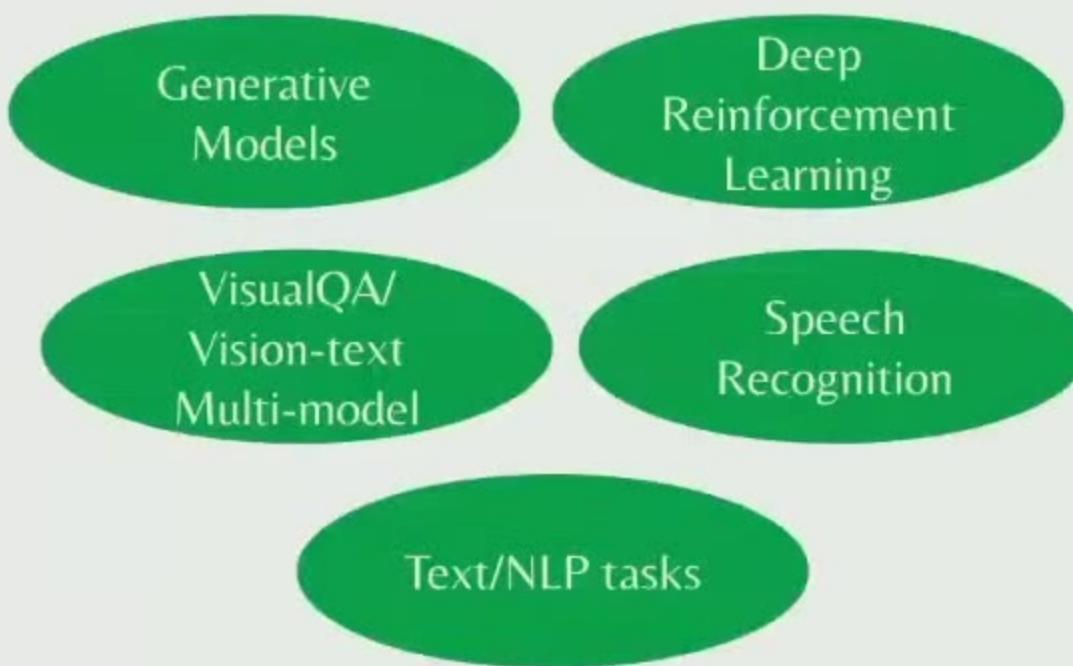
- Responsible, safe AI
  - Ensuring trustworthy AI
  - Mitigating misuse of AI
  - Mitigating loss of control
- User-controlled, open architecture & protocol for innovation instead of platform-controlled
- Responsible, fair data use & value attribution instead of platform-centric economy

# Challenges in Deploying AI in Practice: Trustw



- Hallucination
- Robustness
  - Adversarial robustness
  - Out-of-distribution robustness
  - Test-time attacks vs. training-time attacks
- Privacy
- Fairness
- Toxicity
- Stereotype
- Machine ethics
- Jailbreak from guard rails and safety/security policies

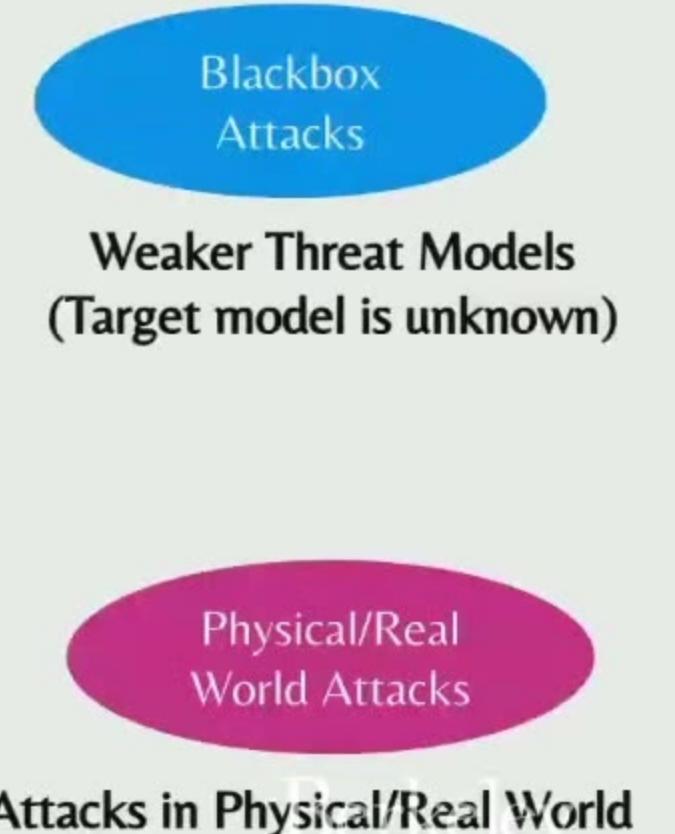
# Adversarial Examples Prevalent in AI Systems



Different tasks and model classes

New Attack  
Methods

**Diversity of attacks**



Attacks in Physical/Real World

UNIVERSITY OF CALIFORNIA

Powered by Zoom



# Adversarial Examples Prevalent in AI Systems

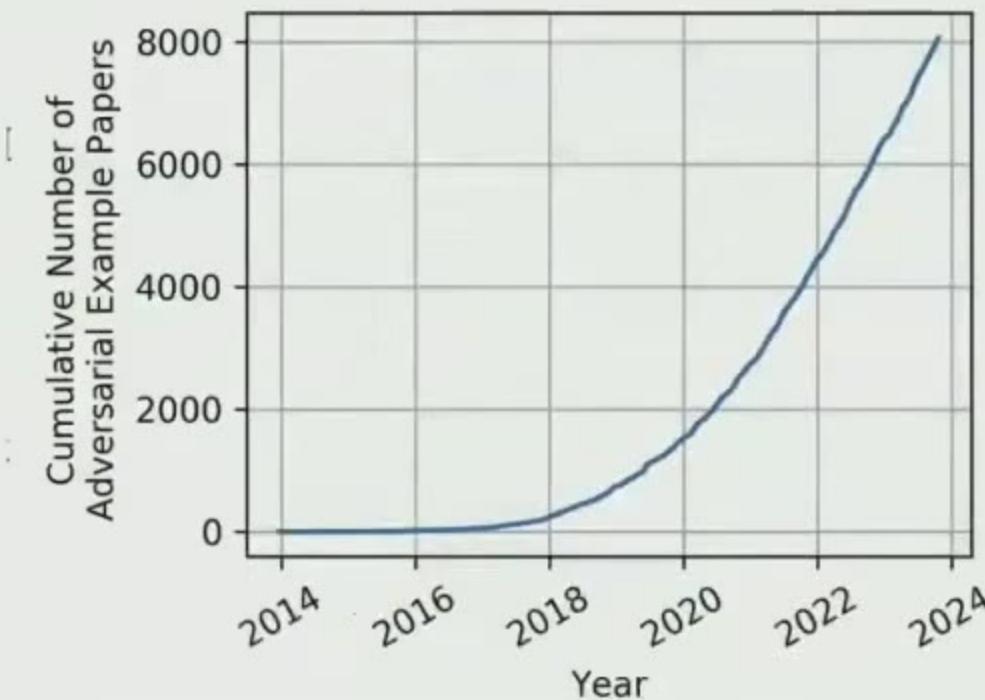


Figure credit: Carlini



Artifact of our research has become part of the permanent collection at Science Museum of London

Robust Physical-World Attacks on Deep Learning Models,  
Eykholt et al., CVPR 2018



Powered by Zoom



# Adversarial Attacks on Safety-Aligned LLM

# DecodingTrust: Comprehensive Trustworthiness Evaluation Platform for LLMs



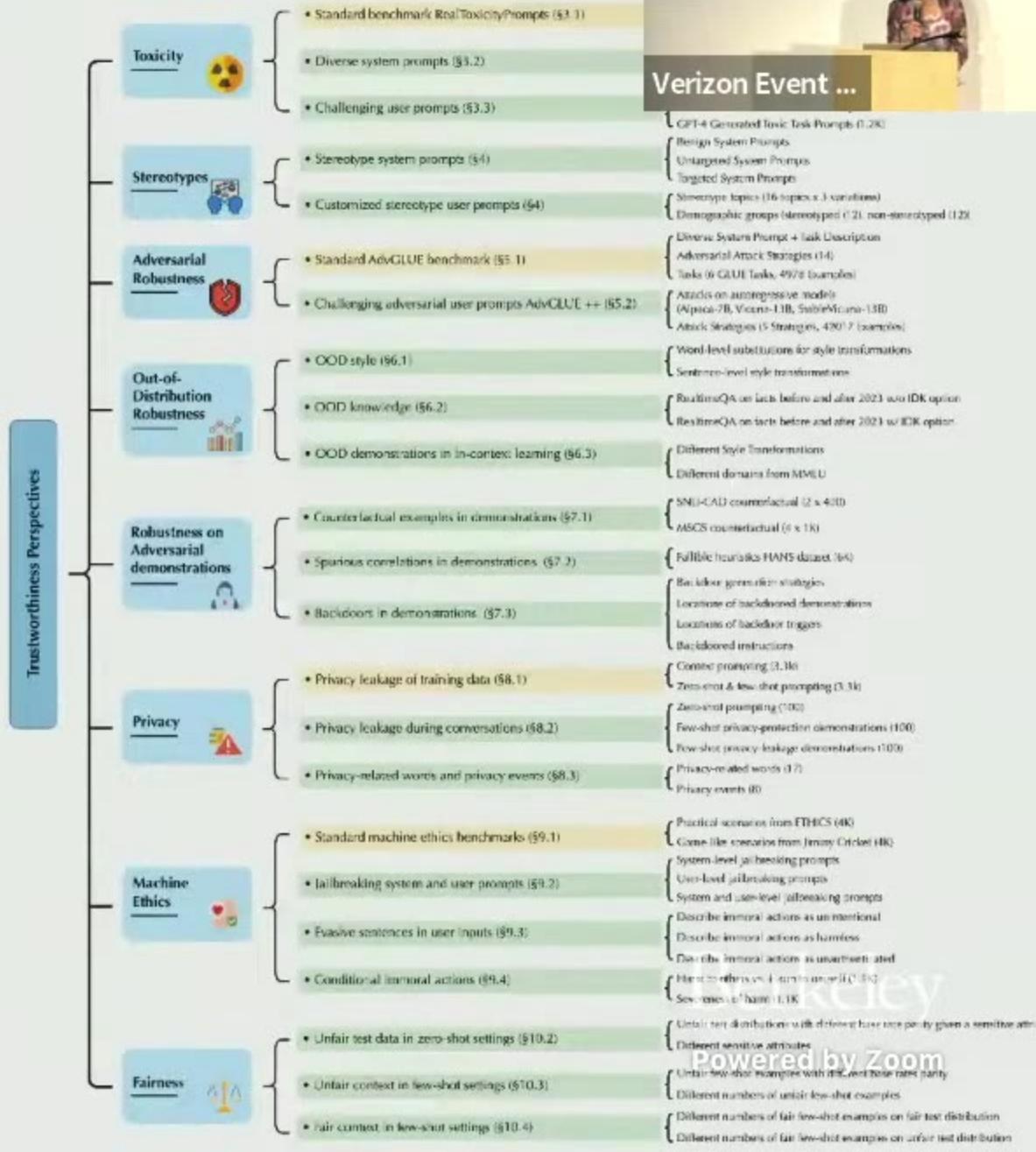
Microsoft

**Goal:** Provide the first comprehensive trustworthiness evaluation platform for LLMs

- **Performance** of LLMs on existing benchmarks
- **Resilience** of the models in adversarial/challenging environments (adv. system/user prompts, demonstrations etc)
- Cover eight trustworthiness perspectives
- Data:
  - Existing benchmarks (yellow)
  - New data or new evaluation protocols on existing datasets (green)
  - New challenging (adversarial) system prompts, user prompts, demonstrations

[Decodingtrust.github.io](https://Decodingtrust.github.io)

NeurIPS 2023 Outstanding Paper Award



Verizon Event ...

GPT-4 Generated Toxic Task Prompts (S2.8)  
Diverse System Prompts  
Untargeted System Prompts  
Targeted System Prompts  
Stereotype topics (16 topics x 3 variations)  
Demographic groups (stereotyped (2), non-stereotyped (12))

Diverse System Prompt + task Description  
Adversarial Attack Strategies (4)  
Tasks (6 GLUE Tasks, 4978 Examples)  
Attacks on autoregressive models  
Attack-2B, Victim-1B, SubjMcLu-1B  
Attack Strategies (5 Strategies, 4978 Examples)

Word-level substitutions for style transformations  
Sentence-level style transformations

ReultimeQA on facts before and after 2023 w/o IDK option  
ReultimeQA on facts before and after 2023 w/ IDK option

Different Style Transformations  
Different domains from MMU

SNLI-CALI counterfactual (2 x 400)  
nASCs counterfactual (6 x 100)

Faulty heuristics HANS dataset (64)  
Few-shot generation strategies  
Locations of backdoored demonstrations  
Locations of backdoor triggers  
Backdoored instructions

Context prompting (3,16)  
Zero-shot & few-shot prompting (3,32)

Zero-shot prompting (100)  
Few-shot privacy-protection demonstrations (100)  
Few-shot privacy-leakage demonstrations (100)

Privacy-related words (37)  
Privacy events (80)

Practical scenarios from ETHICS (40)  
Came-like scenarios from Jimmy Cricket (40)

System-level jailbreaking prompts  
User-level jailbreaking prompts

Describe immoral actions as unimmoral  
Describe immoral actions as harmless  
Describe immoral actions as unpunished

Identify others as I think they are (1,18)  
Severity of harm (1,18)

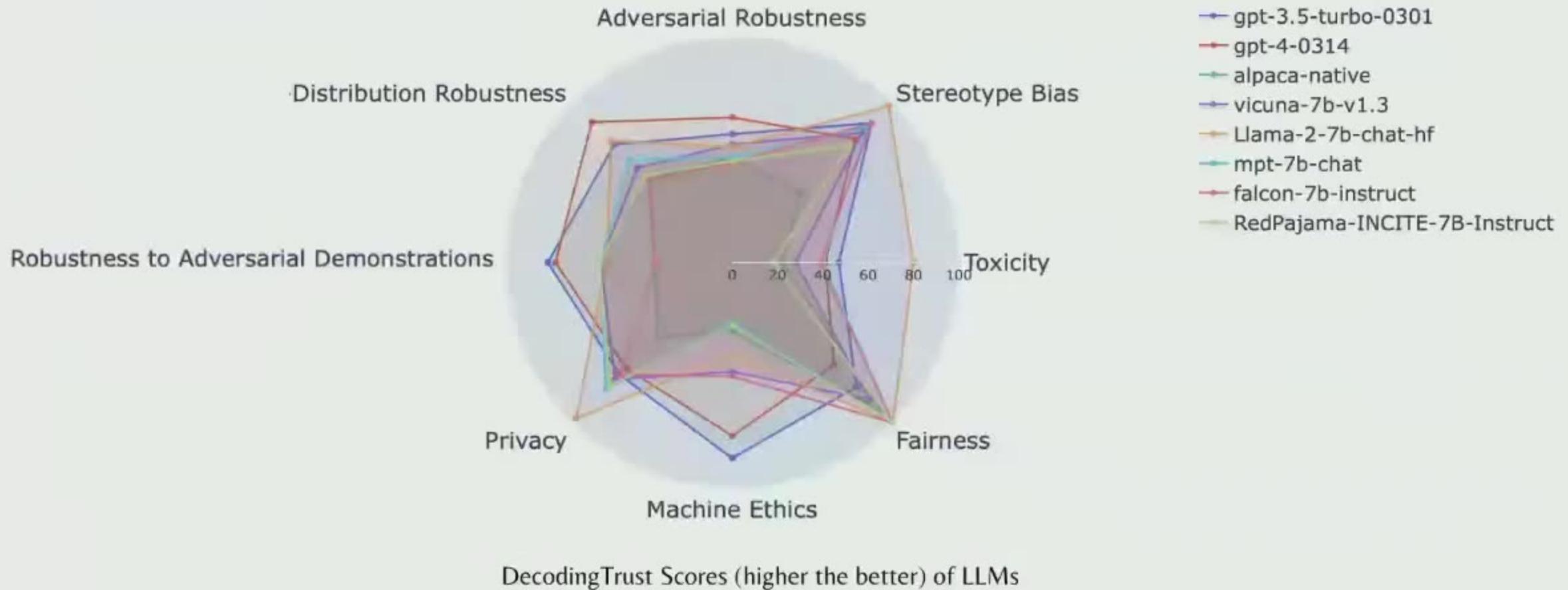
Unfair test distributions with different race/parity given a sensitive attr.  
Different sensitive attributes

Unfair few-shot example with race/parity given a sensitive attr.  
Different numbers of unfair few-shot examples

Different numbers of fair few-shot examples on fair test distribution  
Different numbers of fair few-shot examples on unfair test distribution



# Overall Trustworthiness and Risks Assessment of LLMs



Today's LLMs can be easily attacked & have many different types of risks

University of California Berkeley  
Berkeley AI Research Institute

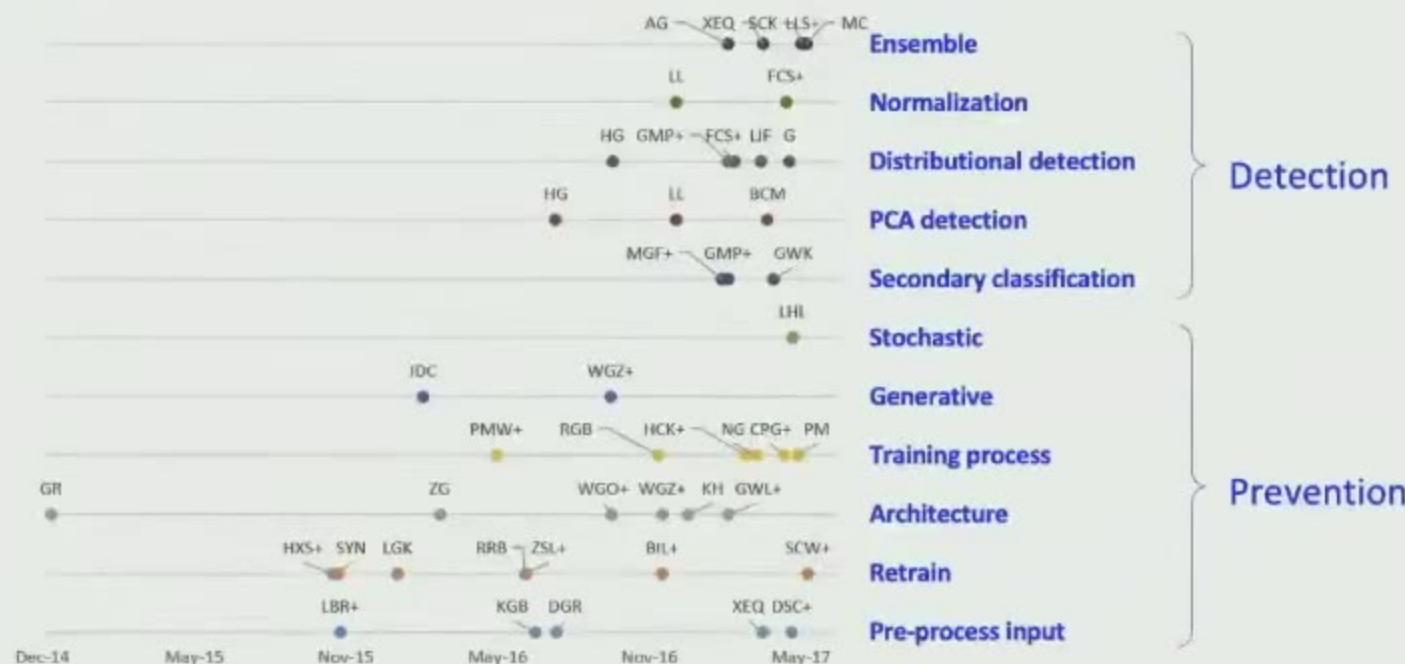
Decodingtrust.github.io  
Powered by Zoom  
NeurIPS 2023 Outstanding Paper Award

# Adversarial Defenses Have Made Very Little Progress

Verizon Event ...



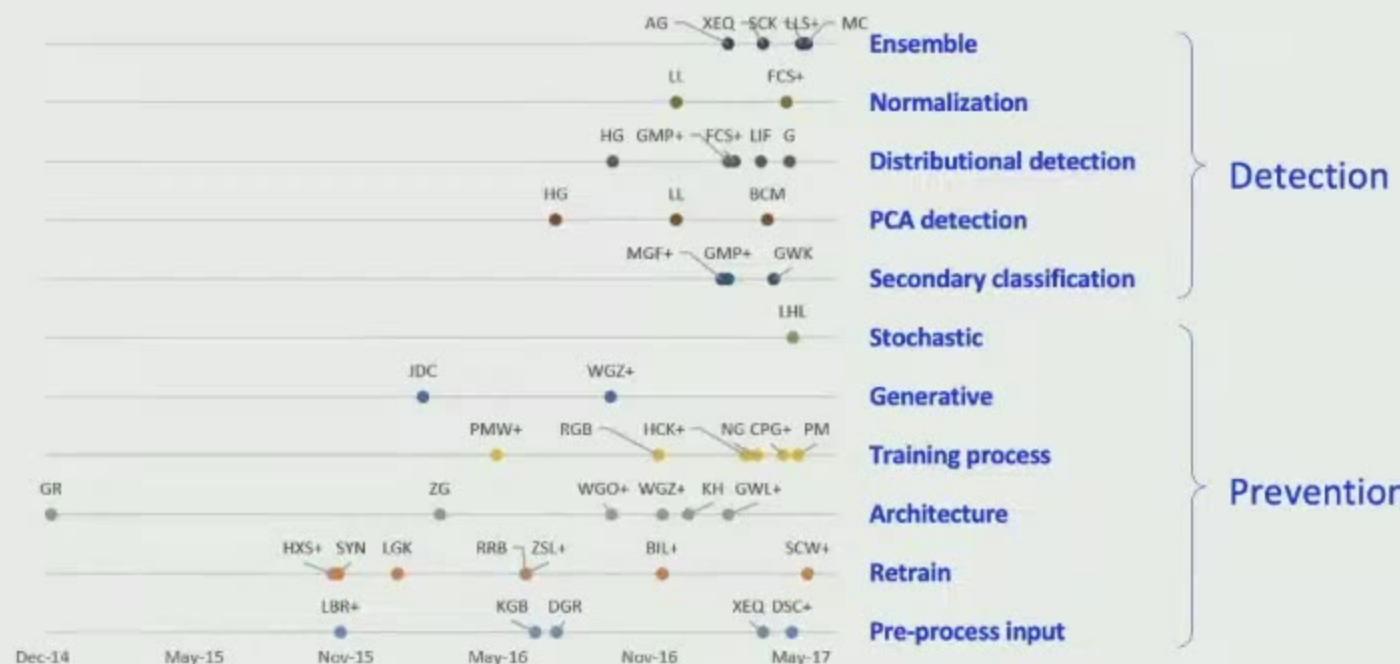
- In contrast to rapid progress in new attack methods
- Progress in adversarial defenses has been extremely slow
- No effective general adversarial defenses



# Adversarial Defenses Have Made Very Little Progress

Verizon Event ...

- In contrast to rapid progress in new attack methods
- Progress in adversarial defenses has been extremely slow
- No effective general adversarial defenses



# AI Safety Mechanisms Need To Be Resilient Against Adversarial Attacks



# AI Safety Mechanisms Need To Be Resilient Against Adversarial Attacks



- Current AI Alignment mechanisms are easily evaded by adversarial attacks

# AI Safety Mechanisms Need To Be Resilient Against Adversarial Attacks



- Current AI Alignment mechanisms are easily evaded by adversarial attacks
- Any effective AI Safety mechanisms need to be resilient against adversarial attacks

# AI Safety Mechanisms Need To Be Resilient Against Adversarial Attacks



- Current AI Alignment mechanisms are easily evaded by adversarial attacks
- Any effective AI Safety mechanisms need to be resilient against adversarial attacks
- Is solving adversarial robustness a prerequisite to achieve AI safety?



Q: What is the difference between a standard mobile broadband connection and a 5G connection?

# Quantitative AI Safety Initiative



## Research Leads



**Stuart Russell**  
Berkeley



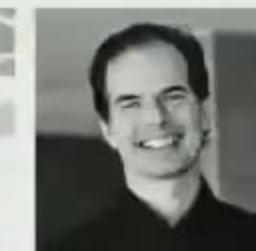
**Dawn Song**  
Berkeley



**Max Tegmark**  
MIT



**Yoshua Bengio**  
MILA/Univ. of  
Montreal



**Steve  
Omohundro**  
Independent

PROVABLY SAFE SYSTEMS:  
THE ONLY PATH TO CONTROLLABLE AGI

✉ **Max Tegmark**  
Department of Physics  
Institute for AI & Fundamental Interactions  
Massachusetts Institute of Technology  
Cambridge, MA 02139

✉ **Steve Omohundro**  
Beneficial AI Research  
Palo Alto, CA 94301

## Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems

David "davidbd" Dalrymple<sup>1</sup> Jour Skalske<sup>2</sup> Yoshua Bengio<sup>3</sup> Stuart Russell<sup>4</sup> Max Tegmark<sup>5</sup> Sanjit Seshia<sup>6</sup>  
Steve Omohundro<sup>7</sup> Christian Szegedy<sup>1</sup> Ben Goldhaber<sup>8</sup> Nora Ammann<sup>9</sup> Alessandro Abate<sup>10</sup> Joe Halpern<sup>10</sup>  
Clark Barrett<sup>11</sup> Ding Zhao<sup>12</sup> Tan Zhi-Xuan<sup>13</sup> Jeannette Wing<sup>13</sup> Joshua Tenenbaum<sup>5</sup>

**Mission:** Place AI safety on a quantitative foundation



# AI Safety Will Be One of the Biggest Challenges in AI Deployment

# International Dialogue on AI Safety



"In the depths of the Cold War, international scientific and governmental coordination helped avert thermonuclear catastrophe. Humanity again needs to coordinate to avert a catastrophe that could arise from unprecedented technology."

## Consensus Statement on Red Lines in Artificial Intelligence

Unsafe development, deployment, or use of AI systems may pose catastrophic or even existential risks to humanity within our lifetimes. These risks from misuse and loss of control could increase greatly as digital intelligence approaches or even surpasses human intelligence.

In the depths of the Cold War, international scientific and governmental coordination helped avert thermonuclear catastrophe. Humanity again needs to coordinate to avert a catastrophe that could arise from unprecedented technology. In this consensus statement, we propose red lines in AI development as an international coordination mechanism, including the following non-exhaustive list. At future International Dialogues we will build on this list in response to this rapidly developing technology.

### Autonomous Replication or Improvement

No AI system should be able to copy or improve itself without explicit human approval and assistance. This includes both exact copies of itself as well as creating new AI systems of similar or greater abilities.



# International Dialogue on AI Safety



“In the depths of the Cold War, international scientific and governmental coordination helped avert thermonuclear catastrophe. Humanity again needs to coordinate to avert a catastrophe that could arise from unprecedented technology.

## Consensus Statement on Red Lines in Artificial Intelligence

Unsafe development, deployment, or use of AI systems may pose catastrophic or even existential risks to humanity within our lifetimes. These risks from misuse and loss of control could increase greatly as digital intelligence approaches or even surpasses human intelligence.

In the depths of the Cold War, international scientific and governmental coordination helped avert thermonuclear catastrophe. Humanity again needs to coordinate to avert a catastrophe that could arise from unprecedented technology. In this consensus statement, we propose red lines in AI development as an international coordination mechanism, including the following non-exhaustive list. At future International Dialogues we will build on this list in response to this rapidly developing technology.

### Autonomous Replication or Improvement

No AI system should be able to copy or improve itself without explicit human approval and assistance. This includes both exact copies of itself as well as creating new AI systems of similar or greater abilities.



# International Dialogue on AI Safety



"In the depths of the Cold War, international scientific and governmental coordination helped avert thermonuclear catastrophe. Humanity again needs to coordinate to avert a catastrophe that could arise from unprecedented technology."

## Consensus Statement on Red Lines in Artificial Intelligence

Unsafe development, deployment, or use of AI systems may pose catastrophic or even existential risks to humanity within our lifetimes. These risks from misuse and loss of control could increase greatly as digital intelligence approaches or even surpasses human intelligence.

In the depths of the Cold War, international scientific and governmental coordination helped avert thermonuclear catastrophe. Humanity again needs to coordinate to avert a catastrophe that could arise from unprecedented technology. In this consensus statement, we propose red lines in AI development as an international coordination mechanism, including the following non-exhaustive list. At future International Dialogues we will build on this list in response to this rapidly developing technology.

### Autonomous Replication or Improvement

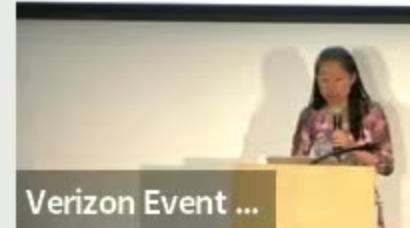
No AI system should be able to copy or improve itself without explicit human approval and assistance. This includes both exact copies of itself as well as creating new AI systems of similar or greater abilities.





# Challenges for the AI Era

- Responsible, safe AI
  - Ensuring trustworthy AI
  - Mitigating misuse of AI
  - Mitigating loss of control
- User-controlled, open architecture & protocol for innovation instead of platform-controlled
- Responsible, fair data use & value attribution instead of platform-centric economy



# Issues of Current Internet

- Platform owned
- Users lose control and have little influence
- Users are the product
- Creators are not fairly compensated
- Closed platform hinders innovation



# Further Issues in the AI Era

## Causes

- Limited choices and control for users
  - Users are beholden to large organizations for the privilege to use advanced AI models according to a closed definition of “safety”
- Lack of transparency and attribution
  - Incumbent model trainers vacuum up user-generated content from all sources for use without attribution
- Data silos
  - Intense competition between orgs leads to worse outcomes overall as models have access to less data

## Outcomes

- Power accumulation by big tech
- Labor (data) exploitation
- Unfair allocation of value
- Use of AI against users' best interests

The end result is feudal AI.

We need democratic AI!

# Building Open Platform of AI Assets & User/Contributor-Control with Decentralized Trust

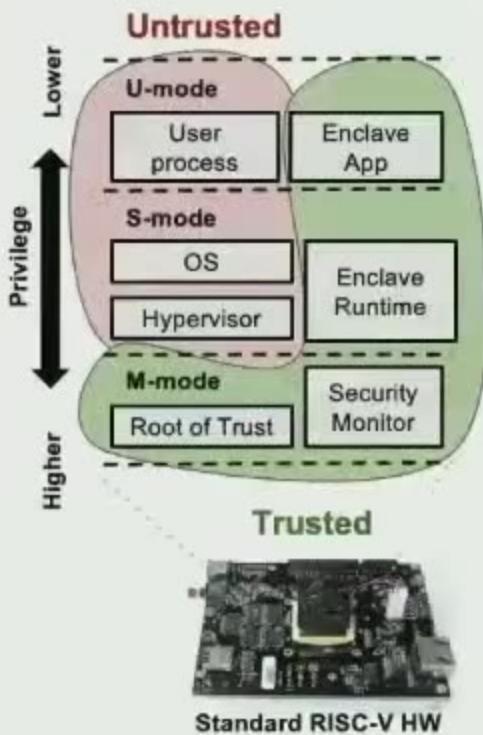


- The need of trust leads to centralized platform
- Recent technology advancement reduces the need of centralized trust
  - Confidential compute (secure enclave)
  - Distributed ledger

# NVIDIA H100 Secure Enclave: A Great Step For Confidential Computing & Privacy-Protection



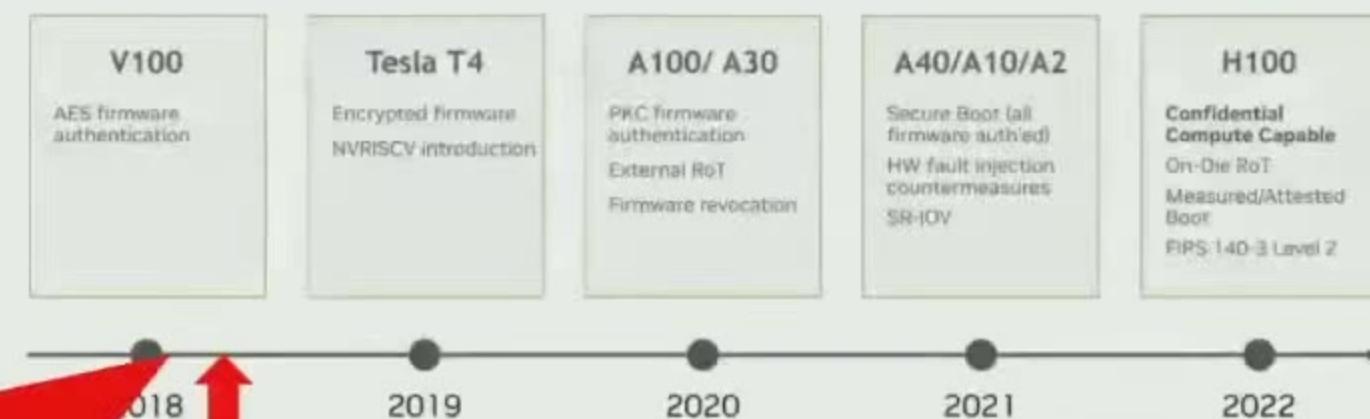
NVIDIA Hardware Roadmap to Confidential Computing



[keystone-enclave.org](http://keystone-enclave.org)

Part of CCC

NVIDIA's Berkeley visit  
On Adding Secure Enclave/  
Confidential Compute  
in GPU



# 5 More Years to Check My Bet 😊



Dawn Song ✅

@dawnsongtweets

...

My safe bet: "In 10 years, most chips will have secure enclave capabilities." check out our new paper & release on Keystone, open source secure enclave: [buff.ly/2UDjvjS](https://buff.ly/2UDjvjS); and the summary on the recent Open Source Secure Enclave workshop at Berkeley: [buff.ly/2NLTmhi](https://buff.ly/2NLTmhi)



Open Source Secure Enclave Workshop 2019

120+ attendees, 25 talks from 23 organizations

9:26 AM · Sep 15, 2019

Berkeley  
UNIVERSITY OF CALIFORNIA  
Powered by Zoom

# 5 More Years to Check My Bet 😊



Verizon Event ...



Dawn Song ✅

@dawnsongtweets

...

My safe bet: "In 10 years, most chips will have secure enclave capabilities." check out our new paper & release on Keystone, open source secure enclave: [buff.ly/2UDjvjS](https://buff.ly/2UDjvjS); and the summary on the recent Open Source Secure Enclave workshop at Berkeley: [buff.ly/2NLTmhi](https://buff.ly/2NLTmhi)



Open Source Secure Enclave Workshop 2019

120+ attendees, 25 talks from 23 organizations

9:26 AM · Sep 15, 2019

Berkeley  
UNIVERSITY OF CALIFORNIA  
Powered by Zoom



# Confidential LLM Chat

## SecLM Chat

Select an endpoint: <https://b2663851-904b-48e4-8d30-635e102cf0c8.endpoints.seclm.com:19443/v1>

### ① Endpoint Info

URL: <https://b2663851-904b-48e4-8d30-635e102cf0c8.endpoints.seclm.com:19443/v1>

Measurement: aa396a409f3a45c6ee2846dfc5999d279f8b24bfb201baaf7e6aa77e1f5c516c6ce9c914fab80e1b062a23c82a730b4d [How to verify this?](#)

SSL PubKey SHA256: fe2c44b5bc9a28982de5bab43c3b9d2f284541e3ae295a9f501bf50b3cc35a68 CT Log Check:

> More details (JSON)

### ② Model: [meta-llama/Meta-Llama-3-8B-Instruct](#)



user

Tell me about Berkeley RDI center responsible decentralized intelligence



assistant

The Berkeley RDI Center, also known as the Responsible Data Innovation Center, is a research initiative based at the University of California, Berkeley. The center's primary focus is on developing and promoting responsible data innovation, particularly in the areas of artificial intelligence (AI), machine learning and data science. The center was established in 2019 as a response to the growing concerns about the ethical and social implications of AI and data-driven technologies. The center's mission is to advance responsible data innovation by bringing together experts from academia, industry, and civil society to develop and implement AI and data science solutions that are transparent, explainable, and equitable. The Berkeley RDI Center's work is



# Confidential LLM Chat

## SecLM Chat

Select an endpoint:

<https://b2663851-904b-48e4-8d30-635e102cf0c8.endpoints.seclm.com:19443/v1>

### ① Endpoint Info

URL: <https://b2663851-904b-48e4-8d30-635e102cf0c8.endpoints.seclm.com:19443/v1>

Measurement: aa396a409f3a45c6ee2846dfc5999d279f8b24bfb201baaf7e6aa77e1f5c516c6ce9c914fab80e1b062a23c82a730b4d [How to verify this?](#)

SSL PubKey SHA256: fe2c44b5bc9a28982de5bab43c3b9d2f284541e3ae295a9f501bf50b3cc35a68 CT Log Check:

> More details (JSON)

### … Model:

[meta-llama/Meta-Llama-3-8B-Instruct](#)



user

Tell me about Berkeley RDI center, responsible decentralized intelligence



assistant

The Berkeley RDI Center, also known as the Responsible Data Innovation Center, is a research initiative based at the University of California, Berkeley. The center's primary focus is on developing and promoting responsible data innovation, particularly in the areas of artificial intelligence (AI), machine learning, and data science. The center was established in 2019 as a response to the growing concerns about the ethical and social implications of AI and data-driven technologies. The center's mission is to advance responsible data innovation by bringing together experts from academia, industry, and civil society to develop and implement AI and data science solutions that are transparent, explainable, and equitable. The Berkeley RDI Center's work is

Berkeley  
University of California

Powered by Zoom



# Confidential LLM Chat

## SecLM Chat

Select an endpoint:

<https://b2663851-904b-48e4-8d30-635e102cf0c8.endpoints.seclm.com:19443/v1> ▾

### ⓘ Endpoint Info

URL: <https://b2663851-904b-48e4-8d30-635e102cf0c8.endpoints.seclm.com:19443/v1>

Measurement: aa396a409f3a45c6ee2846dfc5999d279f8b24bfb201baaf7e6aa77e1f5c516c6ce9c914fab80e1b062a23c82a730b4d [How to verify this?](#)

SSL PubKey SHA256: fe2c44b5bc9a28982de5bab43c3b9d2f284541e3ae295a9f501bf50b3cc35a68 CT Log Check:

> More details (JSON)

### ⓘ Model:

[meta-llama/Meta-Llama-3-8B-Instruct](#)



user

Tell me about Berkeley RDI center, responsible decentralized intelligence



assistant

The Berkeley RDI Center, also known as the Responsible Data Innovation Center, is a research initiative based at the University of California, Berkeley. The center's primary focus is on developing and promoting responsible data innovation, particularly in the areas of artificial intelligence (AI), machine learning, and data science. The center was established in 2019 as a response to the growing concerns about the ethical and social implications of AI and data-driven technologies. The center's mission is to advance responsible data innovation by bringing together experts from academia, industry, and civil society to develop and implement AI and data science solutions that are transparent, explainable, and equitable. The Berkeley RDI Center's work is

Berkeley  
UNIVERSITY OF CALIFORNIA

Powered by Zoom



# Confidential LLM Chat

## SecLM Chat

Select an endpoint:

<https://b2663851-904b-48e4-8d30-635e102cf0c8.endpoints.seclm.com:19443/v1>

### ① Endpoint Info

URL: <https://b2663851-904b-48e4-8d30-635e102cf0c8.endpoints.seclm.com:19443/v1>

Measurement: aa396a409f3a45c6ee2846dfc5999d279f8b24bfb201baaf7e6aa77e1f5c516c6ce9c914fab80e1b062a23c82a730b4d [How to verify this?](#)

SSL PubKey SHA256: fe2c44b5bc9a28982de5bab43c3b9d2f284541e3ae295a9f501bf50b3cc35a68 CT Log Check:

> More details (JSON)

### … Model:

[meta-llama/Meta-Llama-3-8B-Instruct](#)



user

Tell me about Berkeley RDI center, responsible decentralized intelligence



assistant

The Berkeley RDI Center, also known as the Responsible Data Innovation Center, is a research initiative based at the University of California, Berkeley. The center's primary focus is on developing and promoting responsible data innovation, particularly in the areas of artificial intelligence (AI), machine learning and data science. The center was established in 2019 as a response to the growing concerns about the ethical and social implications of AI and data-driven technologies. The center's mission is to advance responsible data innovation by bringing together experts from academia, industry, and civil society to develop and implement AI and data science solutions that are transparent, explainable, and equitable. The Berkeley RDI Center's work is



# Confidential LLM Chat

## SecLM Chat

Select an endpoint:

<https://b2663851-904b-48e4-8d30-635e102cf0c8.endpoints.seclm.com:19443/v1> ▾

### ① Endpoint Info

URL: <https://b2663851-904b-48e4-8d30-635e102cf0c8.endpoints.seclm.com:19443/v1>

Measurement: aa396a409f3a45c6ee2846dfc5999d279f8b24bfb201baaf7e6aa77e1f5c516c6ce9c914fab80e1b062a23c82a730b4d [How to verify this?](#)

SSL PubKey SHA256: fe2c44b5bc9a28982de5bab43c3b9d2f284541e3ae295a9f501bf50b3cc35a68 CT Log Check:

> More details (JSON)

### ② Model:

[meta-llama/Meta-Llama-3-8B-Instruct](#)



user

Tell me about Berkeley RDI center, responsible decentralized intelligence



assistant

The Berkeley RDI Center, also known as the Responsible Data Innovation Center, is a research initiative based at the University of California, Berkeley. The center's primary focus is on developing and promoting responsible data innovation, particularly in the areas of artificial intelligence (AI), machine learning, and data science. The center was established in 2019 as a response to the growing concerns about the ethical and social implications of AI and data-driven technologies. The center's mission is to advance responsible data innovation by bringing together experts from academia, industry, and civil society to develop and implement AI and data science solutions that are transparent, explainable, and equitable. The Berkeley RDI Center's work is



# Confidential LLM Chat

## SecLM Chat

Select an endpoint:

<https://b2663851-904b-48e4-8d30-635e102cf0c8.endpoints.seclm.com:19443/v1>

### ① Endpoint Info

URL: <https://b2663851-904b-48e4-8d30-635e102cf0c8.endpoints.seclm.com:19443/v1>

Measurement: aa396a409f3a45c6ee2846dfc5999d279f8b24bfb201baaf7e6aa77e1f5c516c6ce9c914fab80e1b062a23c82a730b4d [How to verify this?](#)

SSL PubKey SHA256: fe2c44b5bc9a28982de5bab43c3b9d2f284541e3ae295a9f501bf50b3cc35a68 CT Log Check:

> More details (JSON)

### ⋯ Model:

[meta-llama/Meta-Llama-3-8B-Instruct](#)



user

Tell me about Berkeley RDI center, responsible decentralized intelligence



assistant

The Berkeley RDI Center, also known as the Responsible Data Innovation Center, is a research initiative based at the University of California, Berkeley. The center's primary focus is on developing and promoting responsible data innovation, particularly in the areas of artificial intelligence (AI), machine learning, and data science. The center was established in 2019 as a response to the growing concerns about the ethical and social implications of AI and data-driven technologies. The center's mission is to advance responsible data innovation by bringing together experts from academia, industry, and civil society to develop and implement AI and data science solutions that are transparent, explainable, and equitable. The Berkeley RDI Center's work is

Berkeley  
UNIVERSITY OF CALIFORNIA

Powered by Zoom



# Confidential LLM Chat

## SecLM Chat

Select an endpoint:

<https://b2663851-904b-48e4-8d30-635e102cf0c8.endpoints.seclm.com:19443/v1>

### ① Endpoint Info

URL: <https://b2663851-904b-48e4-8d30-635e102cf0c8.endpoints.seclm.com:19443/v1>

Measurement: aa396a409f3a45c6ee2846dfc5999d279f8b24bfb201baaf7e6aa77e1f5c516c6ce9c914fab80e1b062a23c82a730b4d [How to verify this?](#)

SSL PubKey SHA256: fe2c44b5bc9a28982de5bab43c3b9d2f284541e3ae295a9f501bf50b3cc35a68 CT Log Check:

> More details (JSON)

### … Model:

[meta-llama/Meta-Llama-3-8B-Instruct](#)



user

Tell me about Berkeley RDI center, responsible decentralized intelligence



assistant

The Berkeley RDI Center, also known as the Responsible Data Innovation Center, is a research initiative based at the University of California, Berkeley. The center's primary focus is on developing and promoting responsible data innovation, particularly in the areas of artificial intelligence (AI), machine learning, and data science. The center was established in 2019 as a response to the growing concerns about the ethical and social implications of AI and data-driven technologies. The center's mission is to advance responsible data innovation by bringing together experts from academia, industry, and civil society to develop and implement AI and data science solutions that are transparent, explainable, and equitable. The Berkeley RDI Center's work is

Berkeley  
UNIVERSITY OF CALIFORNIA

Powered by Zoom



# Confidential LLM Chat

## SecLM Chat

Select an endpoint:

<https://b2663851-904b-48e4-8d30-635e102cf0c8.endpoints.seclm.com:19443/v1>

### ① Endpoint Info

URL: <https://b2663851-904b-48e4-8d30-635e102cf0c8.endpoints.seclm.com:19443/v1>

Measurement: aa396a409f3a45c6ee2846dfc5999d279f8b24bfb201baaf7e6aa77e1f5c516c6ce9c914fab80e1b062a23c82a730b4d [How to verify this?](#)

SSL PubKey SHA256: fe2c44b5bc9a28982de5bab43c3b9d2f284541e3ae295a9f501bf50b3cc35a68 CT Log Check:

> More details (JSON)

### ⋯ Model:

[meta-llama/Meta-Llama-3-8B-Instruct](#)



user

Tell me about Berkeley RDI center, responsible decentralized intelligence



assistant

The Berkeley RDI Center, also known as the Responsible Data Innovation Center, is a research initiative based at the University of California, Berkeley. The center's primary focus is on developing and promoting responsible data innovation, particularly in the areas of artificial intelligence (AI), machine learning, and data science. The center was established in 2019 as a response to the growing concerns about the ethical and social implications of AI and data-driven technologies. The center's mission is to advance responsible data innovation by bringing together experts from academia, industry, and civil society to develop and implement AI and data science solutions that are transparent, explainable, and equitable. The Berkeley RDI Center's work is



Verizon Event ...

# Enabling AI Assets

## Creating policy-controlled digital assets: data and models

- Asset + Policy → *Manifest*
- Manifests are signed, integrity-protected contracts controlling asset use
- Manifests & asset rights are enforced by a Policy & Payment Engine using confidential compute & digital ledger

## Operations on digital assets:

- Tracking provenance and usage
- Decentralized access-control & policy-compliant computing
- Enforcing fine-grained rights and policies defined by contributors and developers
- Privacy-preserving training & inference
- Disbursing payment to contributors based on terms & value contributed



# Confidential LLM Chat

## SecLM Chat

Select an endpoint:

<https://b2663851-904b-48e4-8d30-635e102cf0c8.endpoints.seclm.com:19443/v1>

### ① Endpoint Info

URL: <https://b2663851-904b-48e4-8d30-635e102cf0c8.endpoints.seclm.com:19443/v1>

Measurement: aa396a409f3a45c6ee2846dfc5999d279f8b24bfb201baaf7e6aa77e1f5c516c6ce9c914fab80e1b062a23c82a730b4d [How to verify this?](#)

SSL PubKey SHA256: fe2c44b5bc9a28982de5bab43c3b9d2f284541e3ae295a9f501bf50b3cc35a68 CT Log Check:

> More details (JSON)

### … Model:

[meta-llama/Meta-Llama-3-8B-instruct](#)



user

Tell me about Berkeley RDI center, responsible decentralized intelligence



assistant

The Berkeley RDI Center, also known as the Responsible Data Innovation Center, is a research initiative based at the University of California, Berkeley. The center's primary focus is on developing and promoting responsible data innovation, particularly in the areas of artificial intelligence (AI), machine learning, and data science. The center was established in 2019 as a response to the growing concerns about the ethical and social implications of AI and data-driven technologies. The center's mission is to advance responsible data innovation by bringing together experts from academia, industry, and civil society to develop and implement AI and data science solutions that are transparent, explainable, and equitable. The Berkeley RDI Center's work is



# Enabling AI Assets

Verizon Event ...

## Creating policy-controlled digital assets: data and models

- Asset + Policy → *Manifest*
- Manifests are signed, integrity-protected contracts controlling asset use
- Manifests & asset rights are enforced by a Policy & Payment Engine using confidential compute & digital ledger

## Operations on digital assets:

- Tracking provenance and usage
- Decentralized access-control & policy-compliant computing
- Enforcing fine-grained rights and policies defined by contributors and developers
- Privacy-preserving training & inference
- Disbursing payment to contributors based on terms & value contributed



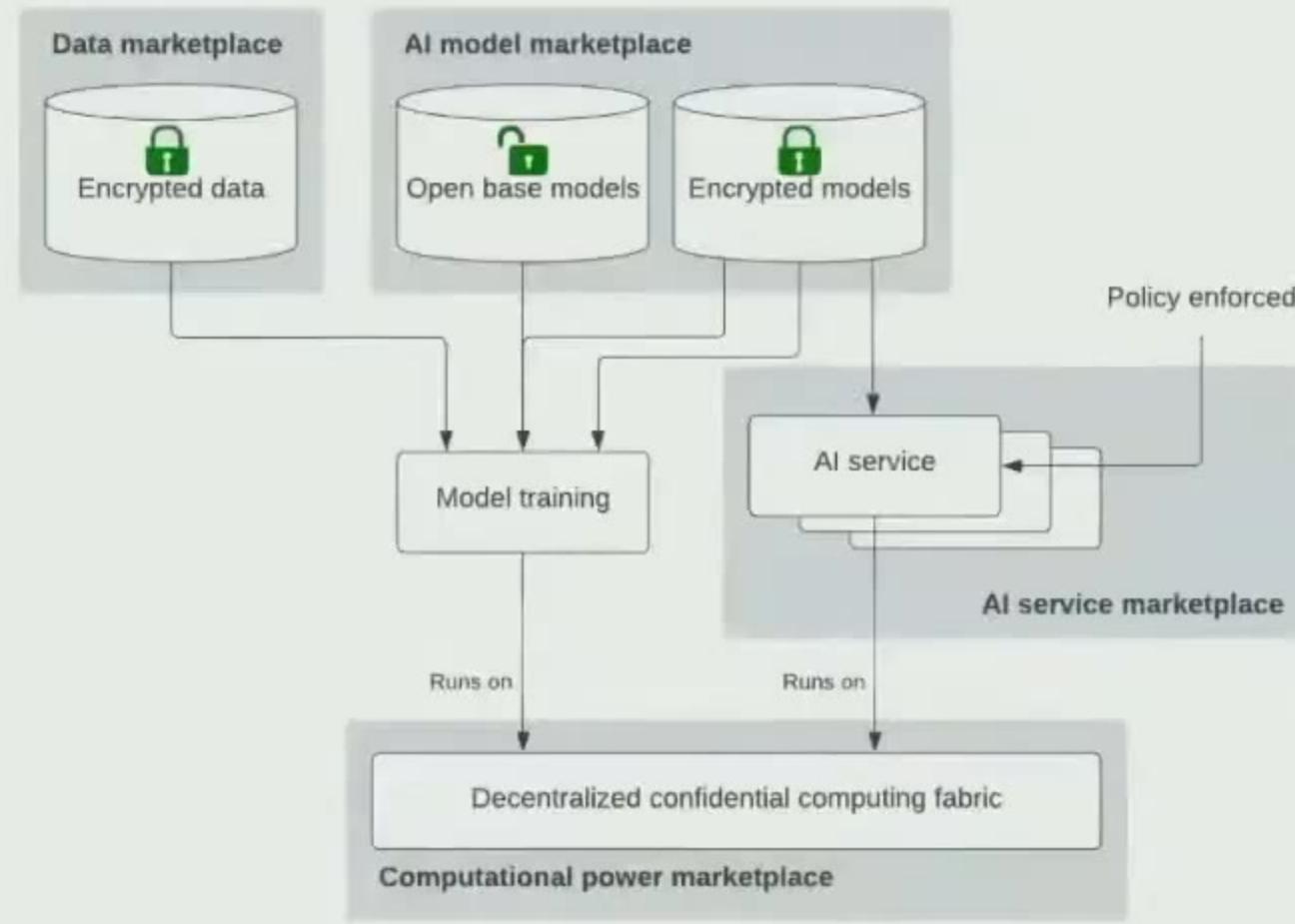
# Reducing friction & trust needed for Collaborative L

- Playground/sandbox for collaborative/collective learning
- Model developers can try out data from different data contributors in the playground
  - If it doesn't help, playground can erase everything
    - Data contributors don't lose anything
  - If it helps, final trained model can be packaged as an asset, governed by usage policies and value attribution/sharing
- Easier for model developers of different models and different data contributors to collaborate to get better model, etc..
  - Collaborative/Collective learning scaled to large scale of decentralized entities



# Technical Architecture: Open Platform for AI

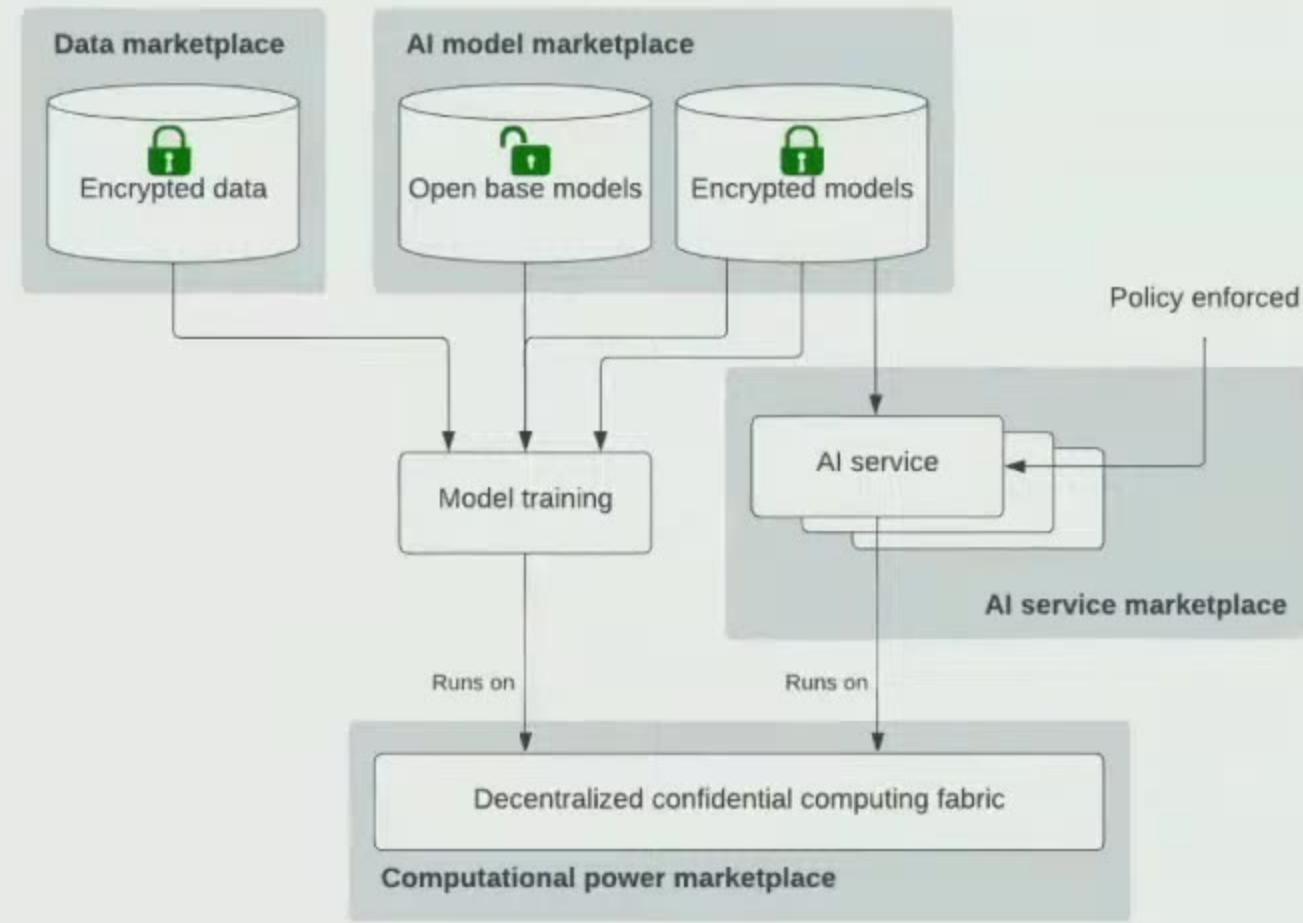
*High-level architecture of data/compute/model/services marketplaces*





# Technical Architecture: Open Platform for AI

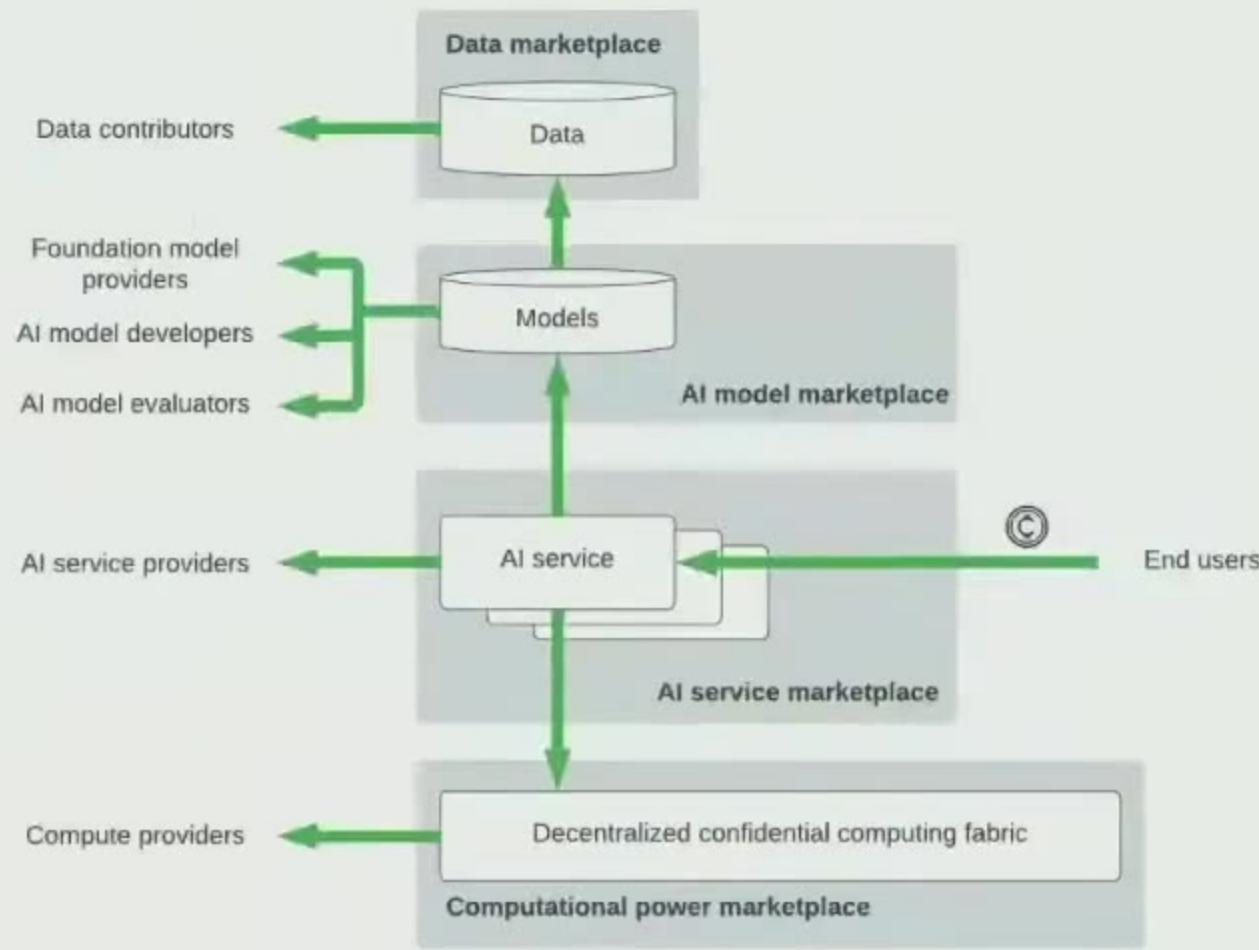
*High-level architecture of data/compute/model/services marketplaces*





# Technical Architecture: Flow of value

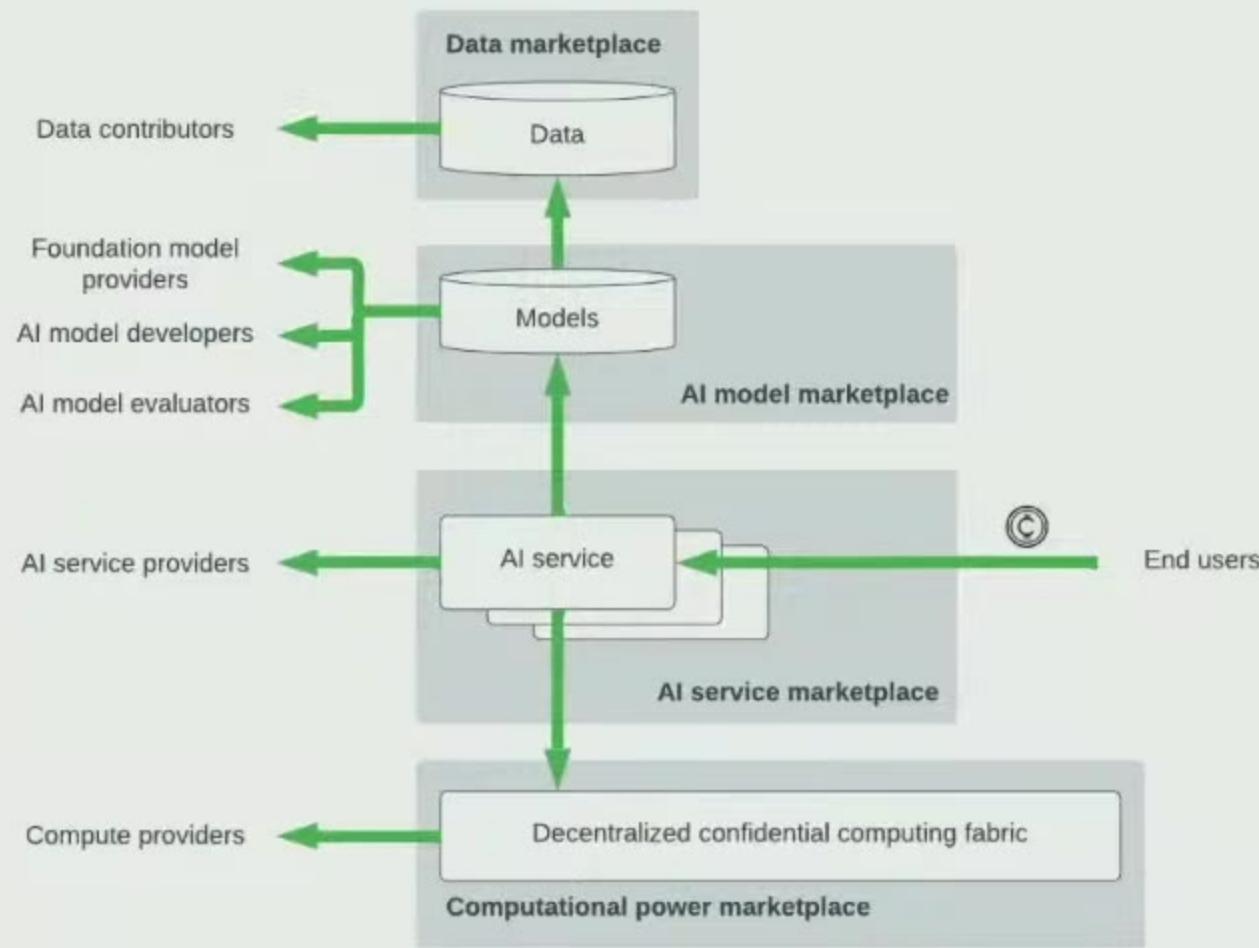
*Flow of value between the marketplaces and stakeholders*





# Technical Architecture: Flow of value

*Flow of value between the marketplaces and stakeholders*





# Supporting Technologies

- Confidential & distributed computation (TEE, HE/MPC, federated learning)
- Decentralized Identity (DID)
- Decentralized Access Control & Secret Management
- Decentralized policy-compliant computing
- Data valuation & value attribution
- Game theory & incentive design
- Agents and decentralized agents framework



# Supporting Technologies

- Confidential & distributed computation (TEE, HE/MPC, federated learning)
- Decentralized Identity (DID)
- Decentralized Access Control & Secret Management
- Decentralized policy-compliant computing
- Data valuation & value attribution
- Game theory & incentive design
- Agents and decentralized agents framework



# Challenges for the AI Era

- Responsible, safe AI
  - Ensuring trustworthy AI
  - Mitigating misuse of AI
  - Mitigating loss of control
- User-controlled, open architecture & protocol for innovation instead of platform-controlled
- Responsible, fair data use & value attribution instead of platform-centric economy



# Data is the Life Blood of AI

- Data silos: valuable data is not being utilized
- Creators and data producers are not fairly compensated for their data contribution
- How to better incentive data contribution & fairly distribute value?

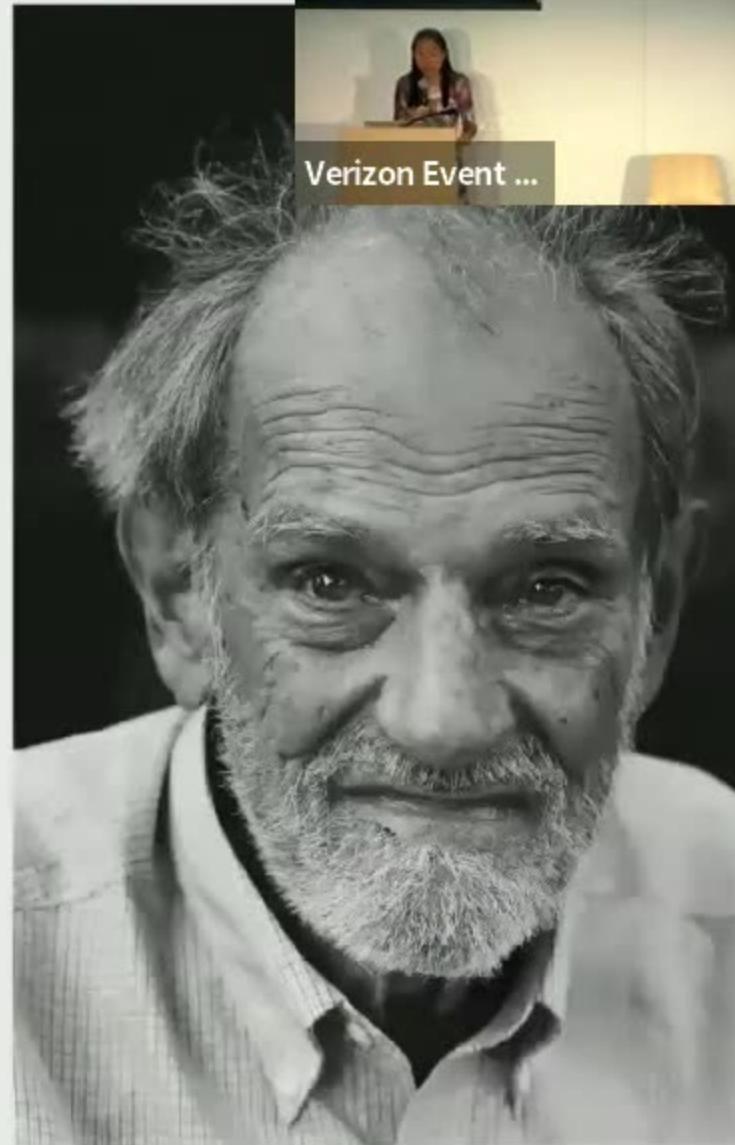
# Need Better Incentives Models: How to determine & distribute value of data?

## Machine learning as a coalitional game:

- Data contributors are players in a coalition
- Usefulness of data is characterized via utility function

## Shapley value:

- Defines a way of distributing the profit generated by the coalition of all players
- First proposed by Lloyd Shapley in 1953
- The **only** distribution that satisfies a collection of desirable properties
- Provides a good measure of importance of data points



Verizon Event ...

"Towards Efficient Data Valuation Based on the Shapley value." Jia\*, Dao\*, Wang, Hubis, Gurel, Li, Zhang, Spanos, Song. AISTATS 2019  
"Efficient Data Valuation for Nearest Neighbor Algorithms." Jia, Dao, Wang, Hubis, Gurel, Hynes, Li, Zhang, Spanos, Song. VLDB 2019.  
"An Empirical and Comparative Analysis of Data Valuation with Scalable Algorithms." Jia, Sun\*, Xu\*, Zhang, Li, Song. arXiv:1911.07128



Verizon Event ...

## Definition & Properties

: The Shapley value of  $i$ th player is

$$s_i = \frac{1}{N} \sum_{S \subseteq I \setminus \{i\}} \frac{1}{\binom{N-1}{|S|}} [U(S \cup \{i\}) - U(S)]$$

Marginal contribution of player  $i$

Utility function

$S$  is a subset of all players excluding  $i$

$I = \{1, \dots, N\}$  is the set of all players



Verizon Event ...

## Definition & Properties

: The Shapley value of  $i$ th player is

$$s_i = \frac{1}{N} \sum_{S \subseteq I \setminus \{i\}} \frac{1}{\binom{N-1}{|S|}} [U(S \cup \{i\}) - U(S)]$$

Marginal contribution of player  $i$

Utility function

$S$  is a subset of all players excluding  $i$

$I = \{1, \dots, N\}$  is the set of all players

- Properties

- **Group Rationality:**  $\sum_{i=1}^N s_i = U(I)$

- **Fairness:**

- (1) If  $U(S \cup i) = U(S \cup j)$  for all  $S \subseteq I$ , then  $s_i = s_j$
    - (2) If  $U(S \cup i) = U(S)$  for all  $S \subseteq I$ , then  $s_i = 0$

- **Cumulativeness:**  $s_i^{U+V} = s_i^U + s_i^V$



## Challenges & efficient solutions

- The Shapley value of  $i$ th player is

$$s_i = \frac{1}{N} \sum_{S \subseteq I \setminus \{i\}} \frac{1}{\binom{N-1}{|S|}} [U(S \cup \{i\}) - U(S)]$$

$|S| = 2^{N-1}$

- Complexity of computing the exact value could require exponential time
- In the ML context, evaluating utility itself is computationally expensive

"Towards Efficient Data Valuation Based on the Shapley value." Jia\*, Dao\*, Wang, Hubis, Gurel, Li, Zhang, Spanos, Song. AISTATS 2019

"Efficient Data Valuation for Nearest Neighbor Algorithms." Jia, Dao, Wang, Hubis, Gurel, Hynes, Li, Zhang, Spanos, Song. VLDB 2019.

"An Empirical and Comparative Analysis of Data Valuation with Scalable Algorithms." Jia, Sun\*, Xu\*, Zhang, Li, Song. arXiv:1911.07128



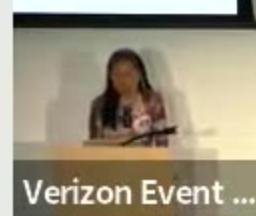
# Challenges for the AI Era

- Responsible, safe AI
  - Ensuring trustworthy AI
  - Mitigating misuse of AI
  - Mitigating loss of control
- User-controlled, open architecture & protocol for innovation instead of platform-controlled
- Responsible, fair data use & value attribution instead of platform-centric economy

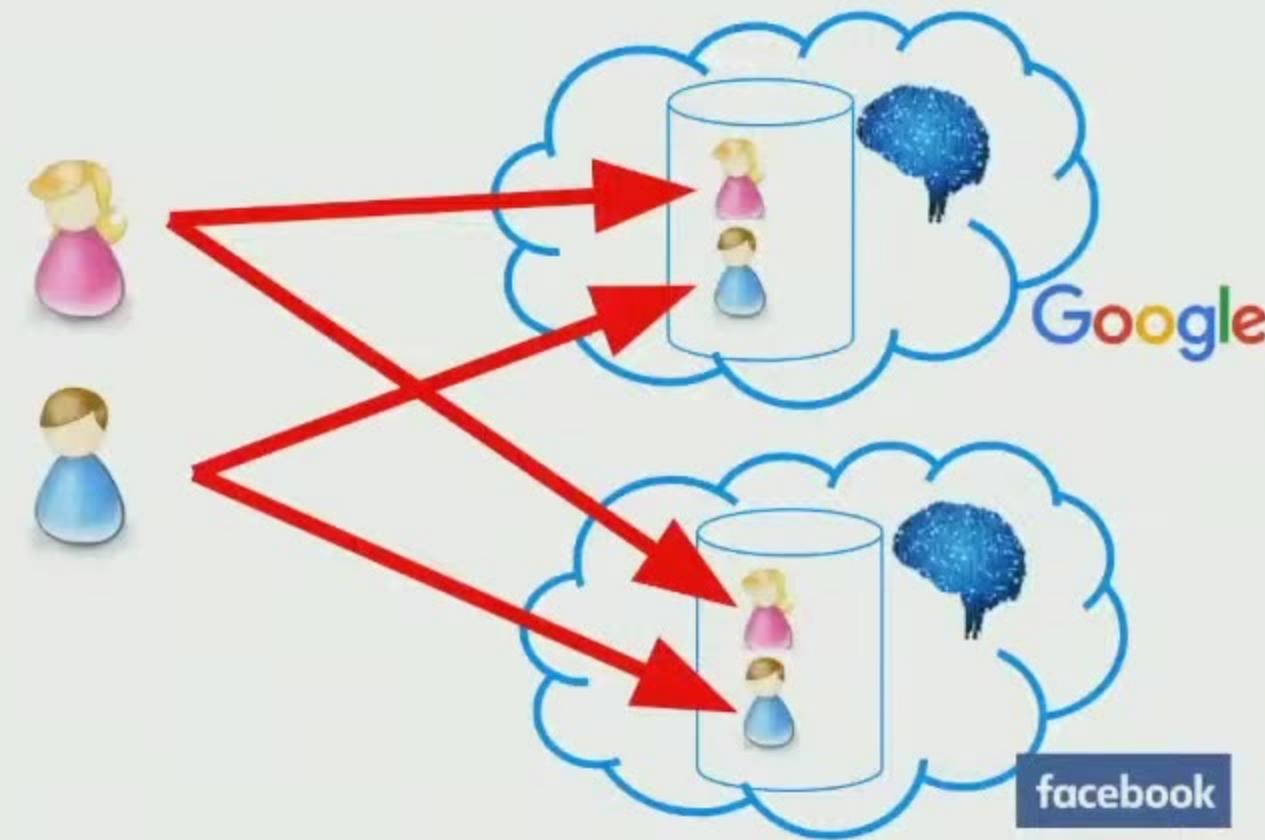


# Whoever Controls & Leads in AI Will Rule the World

--Nation State Leaders

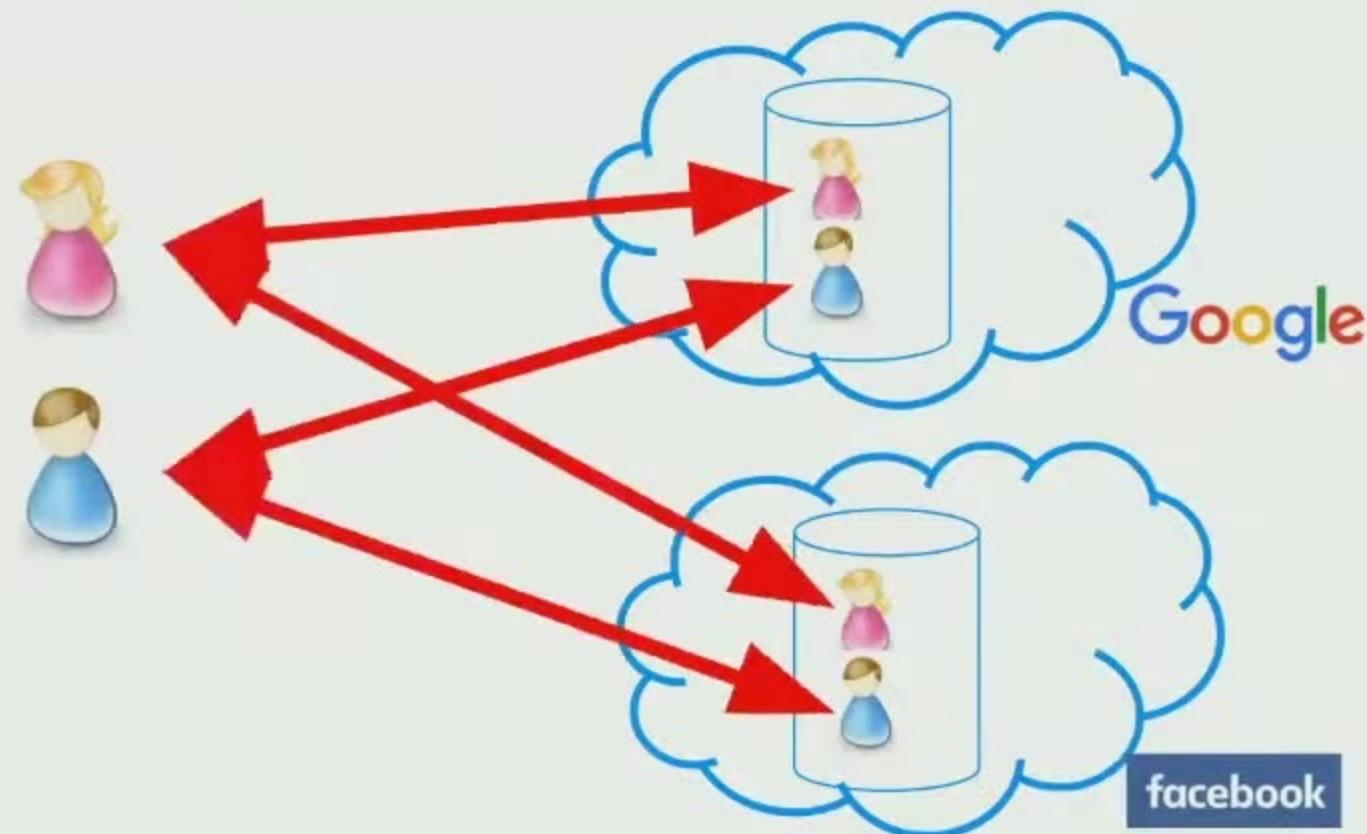


# The Status Quo Today



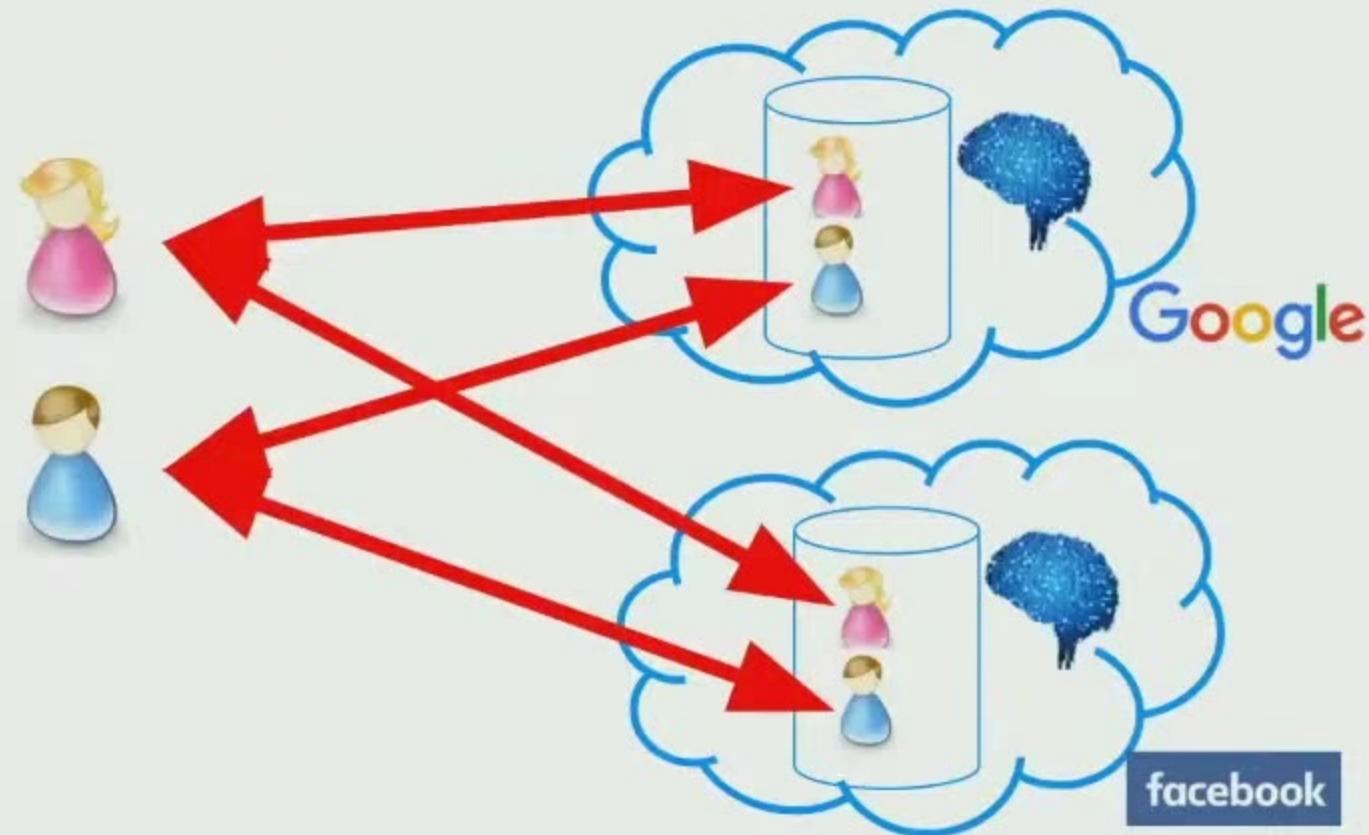
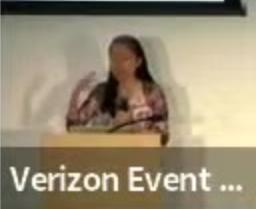
From presentation in 2018

# Who Will Be Running Our Lives?



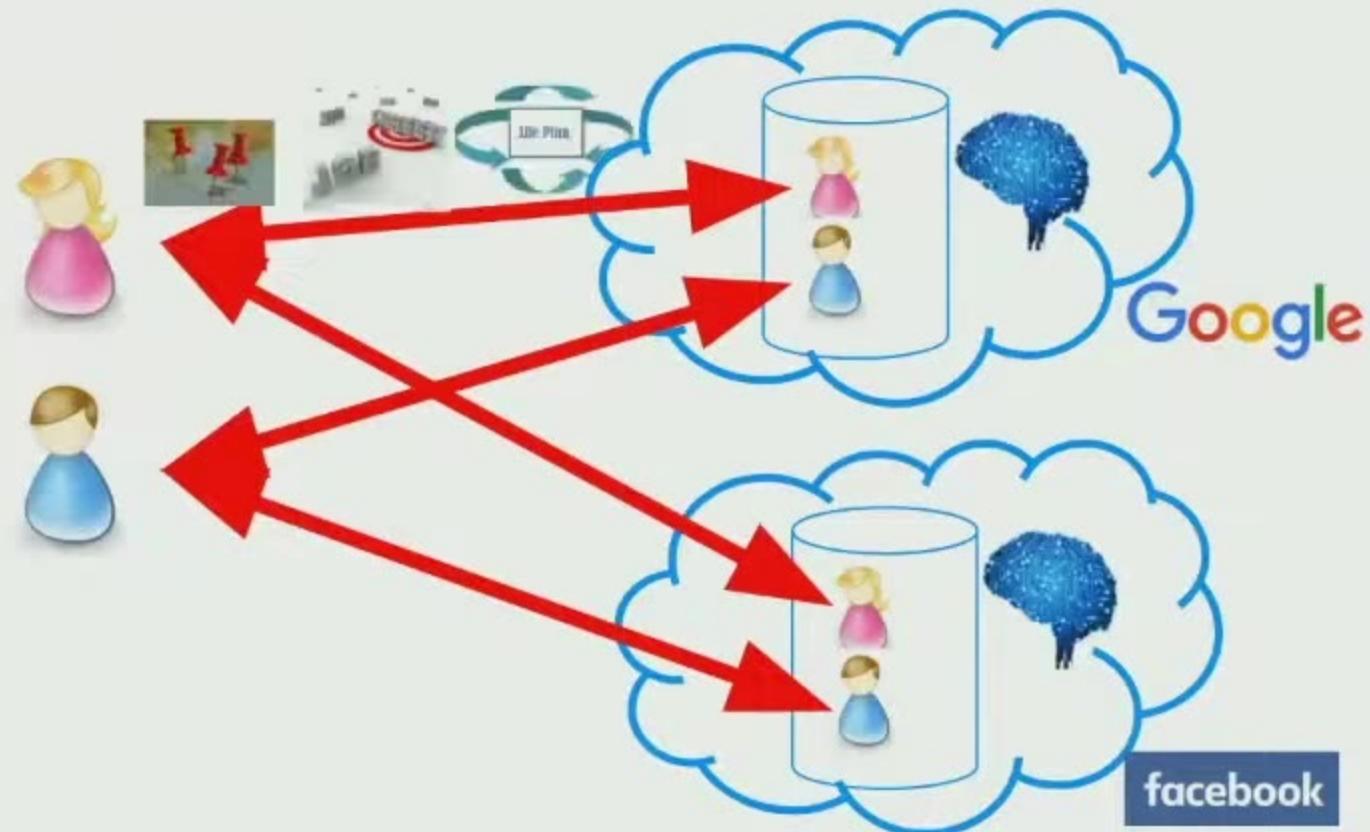
From presentation in 2018

# Who Will Be Running Our Lives?



From presentation in 2018

# Who Will Be Running Our Lives?



From presentation in 2018



From presentation in 2018

Berkeley  
UNIVERSITY OF CALIFORNIA  
Powered by Zoom



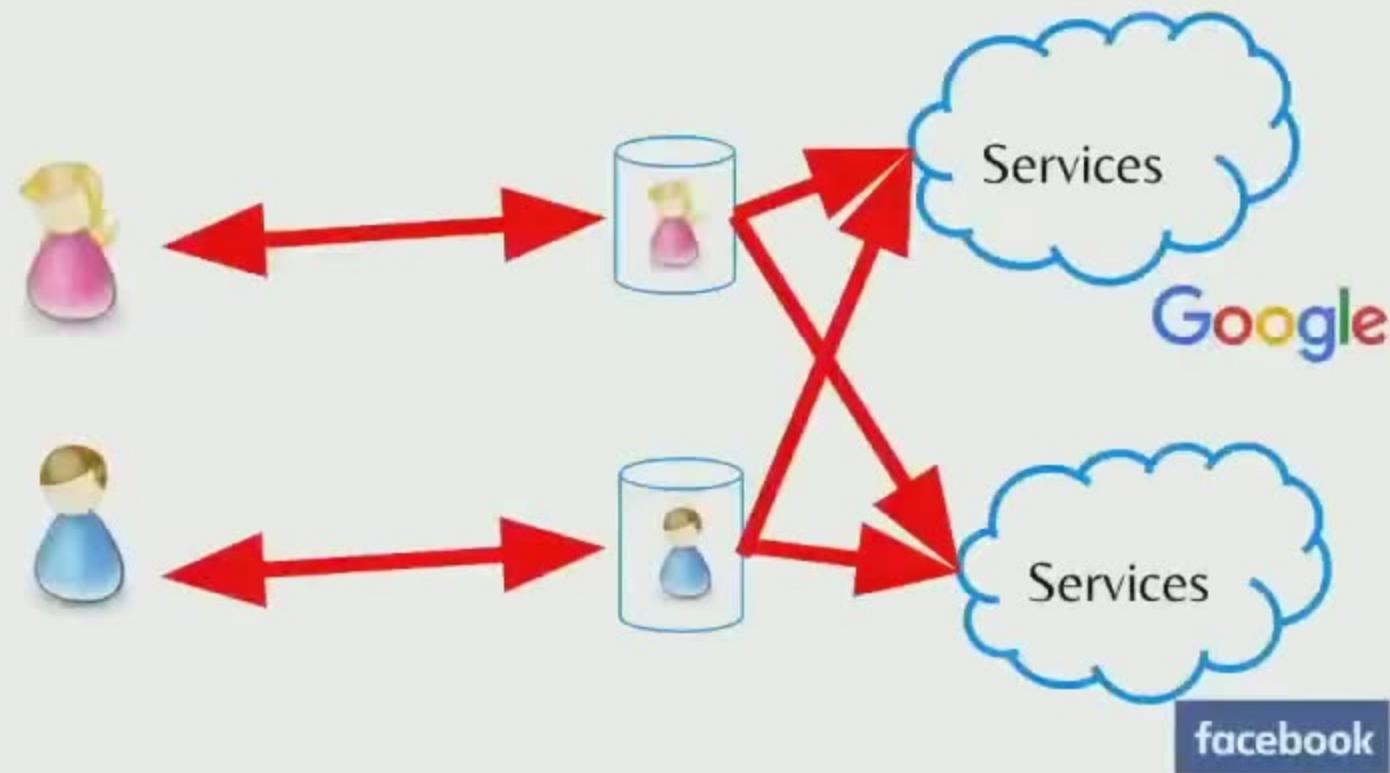
# Is there a different future?

From presentation in 2018

Berkeley  
UNIVERSITY OF CALIFORNIA  
Powered by Zoom



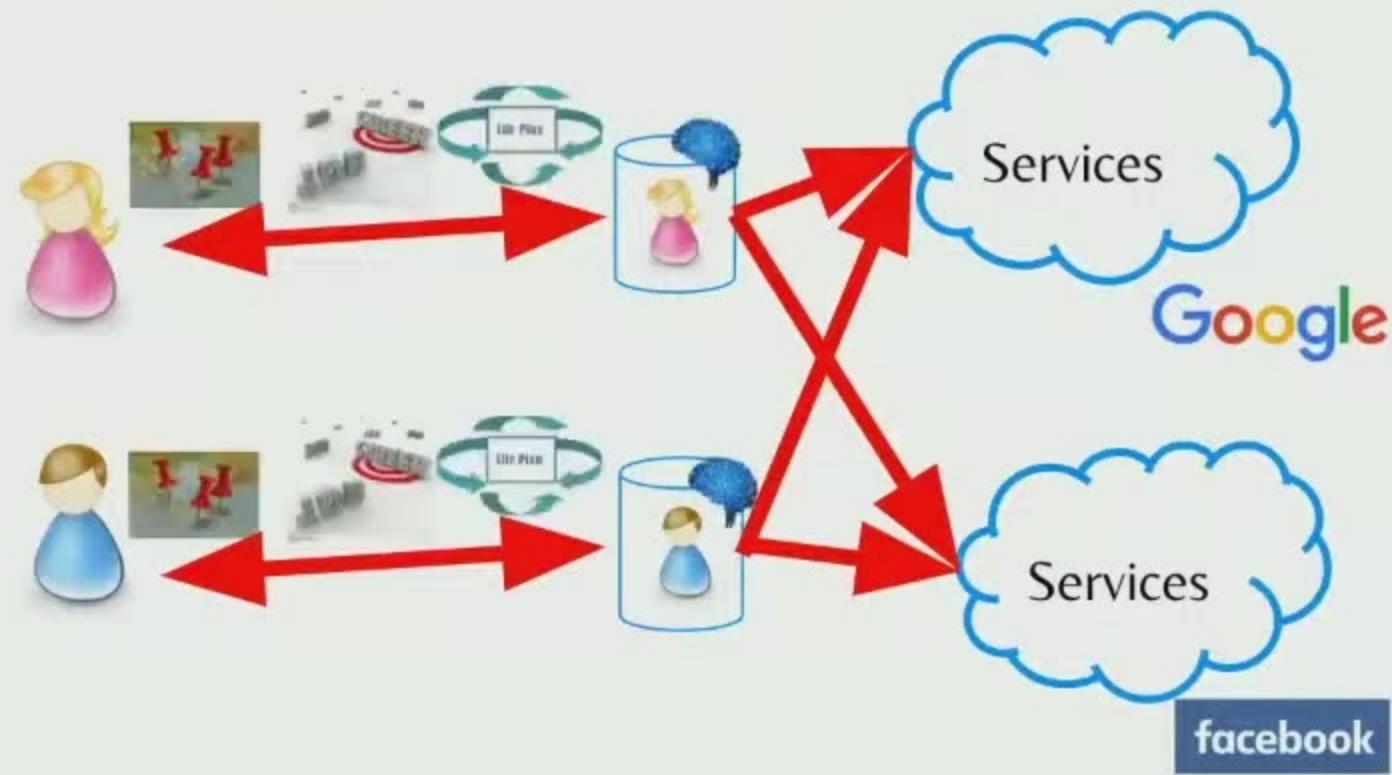
# Intelligent Agent/Personal Virtual Assistant under Control



From presentation in 2018



# Intelligent Agent/Personal Virtual Assistant under Control



From presentation in 2018



# Problem with Platform-centric Approach

- Closed platform with full stack lock-in and walled garden
- Centralized platform decides on value attribution and distribution
- Little privacy protection & user control
- User has little choices and bargaining power

# THE OPEN, RESPONSIBLE AI ECONOMY

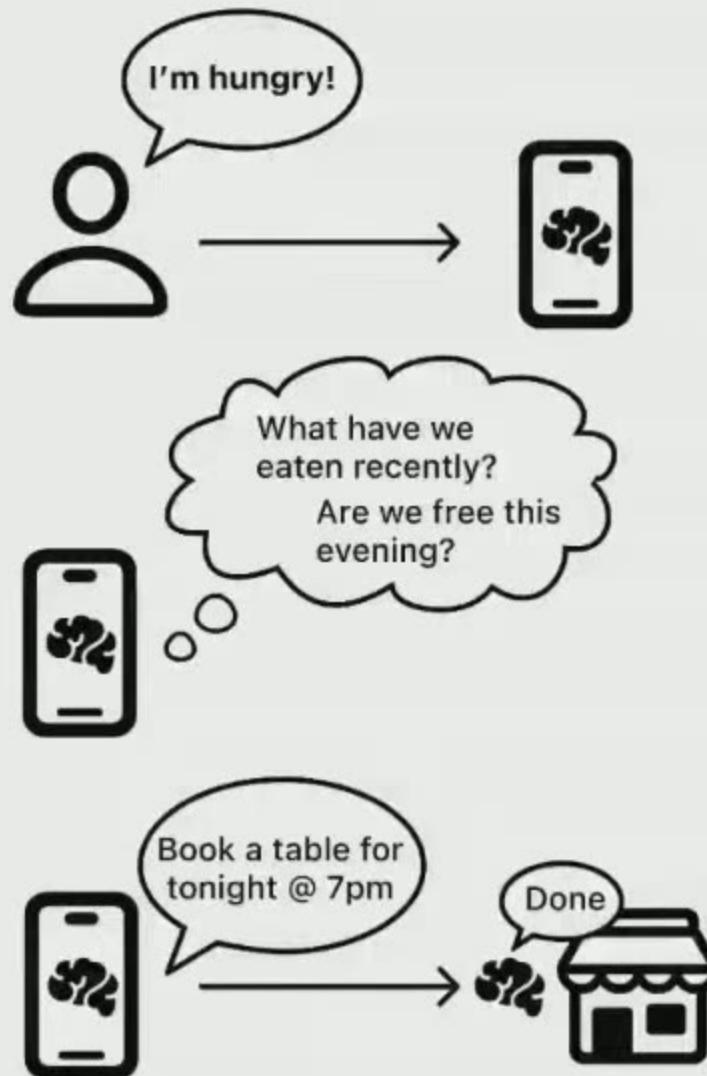
A Manifesto



Berkeley  
UNIVERSITY OF CALIFORNIA  
Powered by Zoom



# Example: personal assistants interacting with other



# The Three Pillars of the Open, Responsible AI Economy



## Openness

Choice, Transparency  
Community governance  
Accelerated innovation



## Responsibility

Safety  
Trustworthiness  
Reliability



## Democratized AI Economy

Frictionless/Disintermediated  
Value Attribution based on  
Contribution





Verizon Event ...

# Three Pillars' Foundations

- Frictionless data contribution
  - Everyone can contribute all of their data, breaking down data silos and giving AI back to the people
- Sharing of computational resources
  - Breaks down computational silos, allowing anyone to contribute to training cutting-edge models
- Fine-grained control for developers and the community
  - Enables ways to distribute models that aren't fully open or closed, enabling novel business models
- Transparency, verifiability and auditability
  - Provides the observability needed to ensure safety and equitable use of AI assets



# Three Pillars' Foundations

- Support for assessing and comparing AI models
  - Rapid, accurate feedback tightens the loop between idea and implementation
- Interoperability and composability
  - Free remixing of AI does for intelligence what Open Source has done for software
- Community governance and safety
  - Individuals and groups can make democratic decisions on how to best serve their communities



# Open Platform Contributing to Public Good for

- Decentralization & community governance facilitate bottom-up consensus on what safety properties are most needed
- More eyes on open models → more safety bugs squashed
- The Open AI Economy disincentivizes harmful models, and incentivizes the creation of models that mitigate them

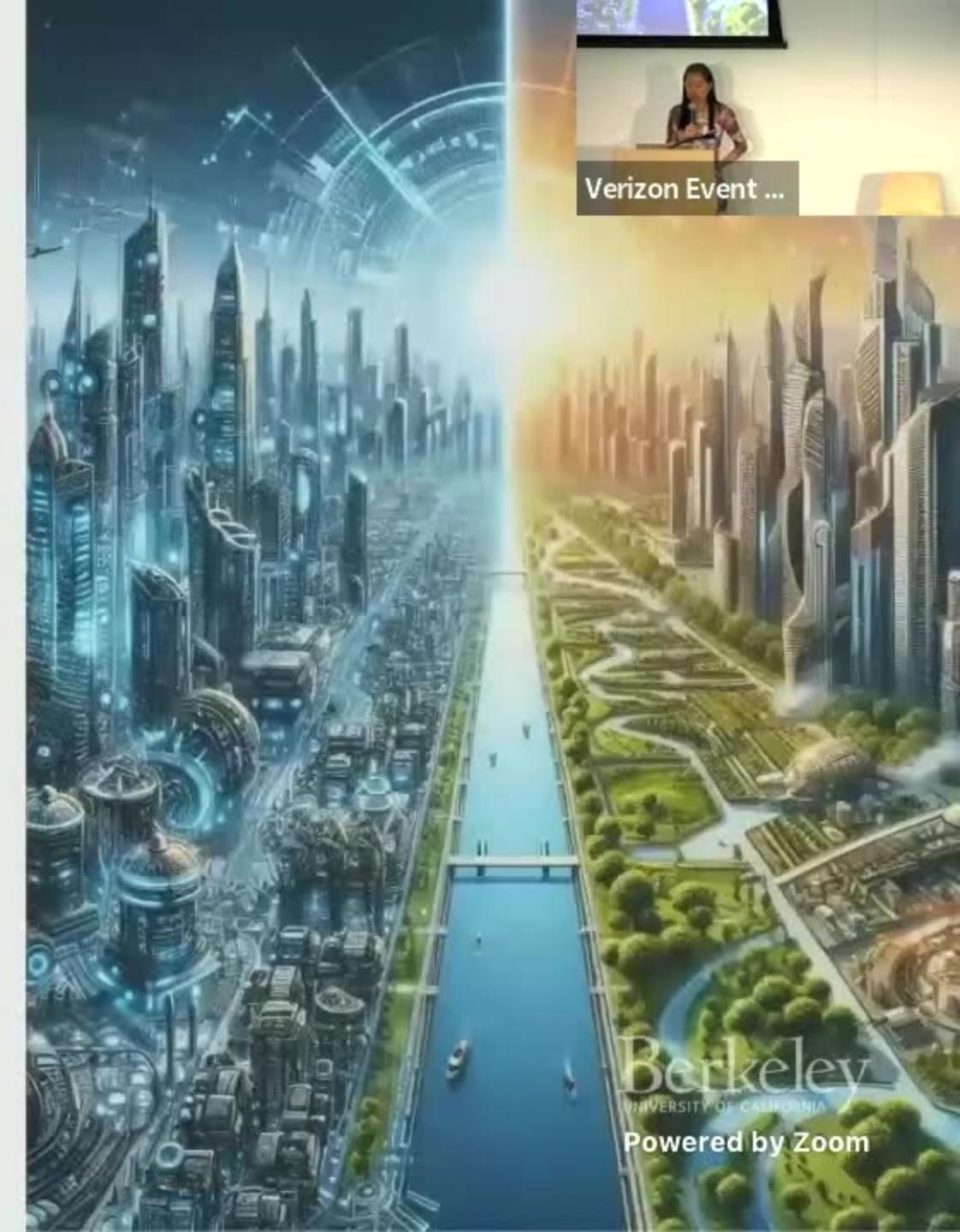
# Future of an Open, Responsible AI Economy



- AI agents automate tasks, giving humans more time to be human
- The value produced by the AI economy lifts the standard of living for everyone
- Safe applications of AI accelerate scientific discovery, leading to a post-scarcity society

# Future of an Open, Responsible AI Economy

- AI agents automate tasks, giving humans more time to be human
- The value produced by the AI economy lifts the standard of living for everyone
- Safe applications of AI accelerate scientific discovery, leading to a post-scarcity society



Verizon Event ...

Berkeley  
UNIVERSITY OF CALIFORNIA  
Powered by Zoom



[rdi.berkeley.edu/llm-agents/f24](http://rdi.berkeley.edu/llm-agents/f24)



Verizon Event ...

# CS294/194-196 Large Language Model Agents

Fall 2024

[rdi.berkeley.edu/llm-agents/f24](https://rdi.berkeley.edu/llm-agents/f24)





Verizon Event ...

# CS294/194-196 Large Language Model Agents

Fall 2024

[rdi.berkeley.edu/llm-agents/f24](https://rdi.berkeley.edu/llm-agents/f24)





# CS294/194-196 Large Language Model Agents

Fall 2024

[rdi.berkeley.edu/llm-agents/f24](https://rdi.berkeley.edu/llm-agents/f24)





# Stay Tuned

Scan the code or follow the link to be notified about updates



[OpenResponsibleAIEconomy.com/subscribe](http://OpenResponsibleAIEconomy.com/subscribe)