Where Is AI Headed?

Dean Wampler, Ph.D.
The AI Alliance and IBM Research
thealliance.ai
deanwampler.com/talks

deanwampler.com/talks

© 2023-2024, Dean Wampler, except where noted. Photos at https://www.flickr.com/photos/deanwampler/

About the Images…

https://discuss.systems/@mikemathia@ioc.exchange/112687372445996049

Outline

- The AI Alliance - Why?
- Should AI Be “Open” or “Closed”?
- The Challenges to Success
- Generative AI in Five Years??

thealliance.ai

The AI Alliance

A community of technology creators, developers and adopters collaborating to advance safe, responsible AI rooted in open innovation.

Diagram as of February.
>110 Now

The **AI Alliance**

thealliance.ai

A community of technology creators, developers and adopters collaborating to advance safe, responsible AI rooted in open innovation.

Six Focus Areas:

1. Education, skills building, and research
2. Trust and safety
3. Tools for building models and applications
4. Hardware portability
5. Open models and datasets
6. Policy and regulations

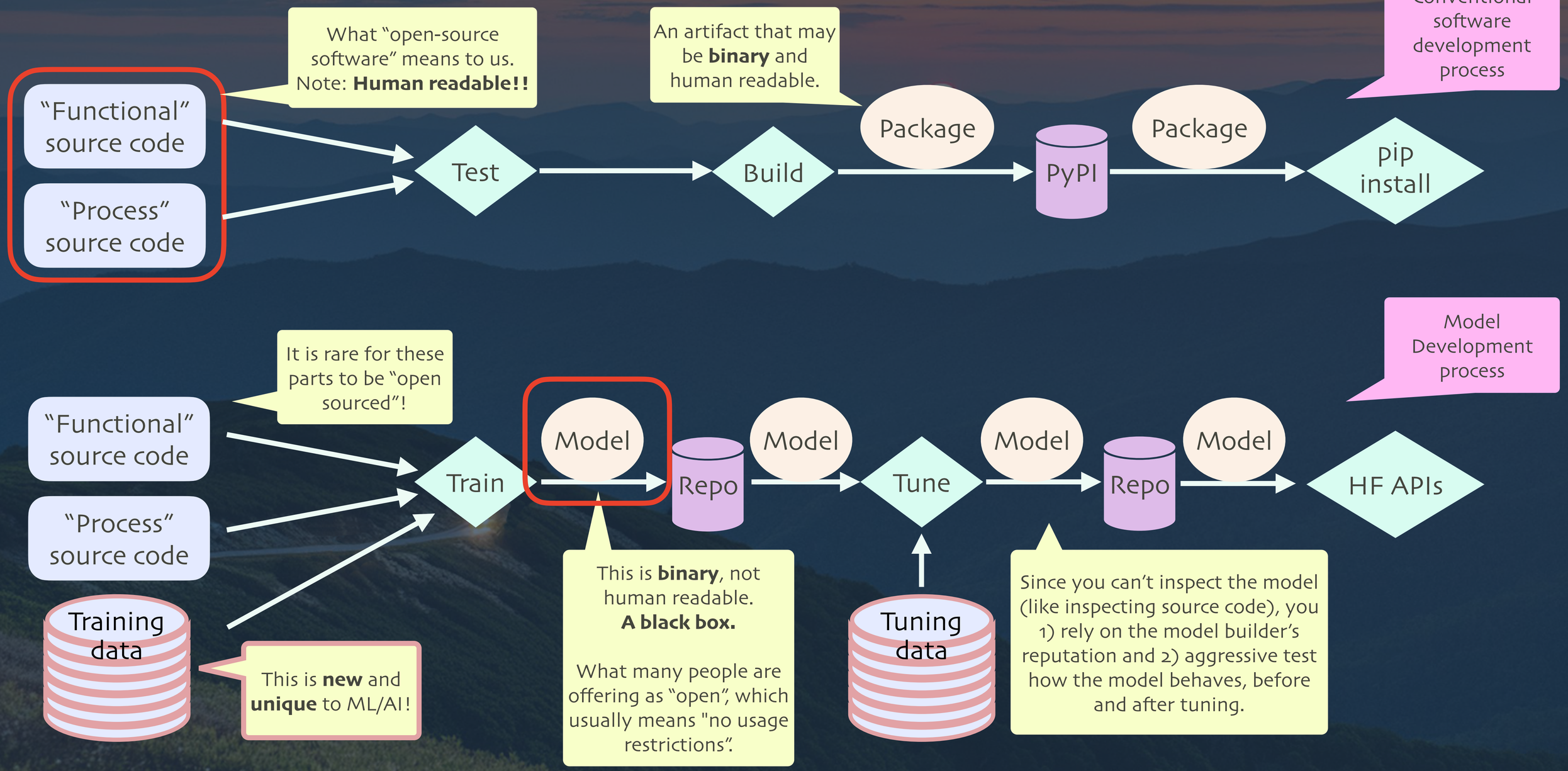
Spreading knowledge, research

Technical initiatives

Maximize access, with safety


Should AI be
“open” or “closed”?

First, what does open mean??



Why Open (as Much as Possible)?

- Free to use as you see fit without undo restrictions.
- Free to innovate in new ways.
- Easier to inspect for bugs, security flaws.
  - For data, easier to inspect for "bad" data.
    - E.g., hate speech, copyrighted content, etc.

Do we want AI technology controlled by a few entities or more widely available?

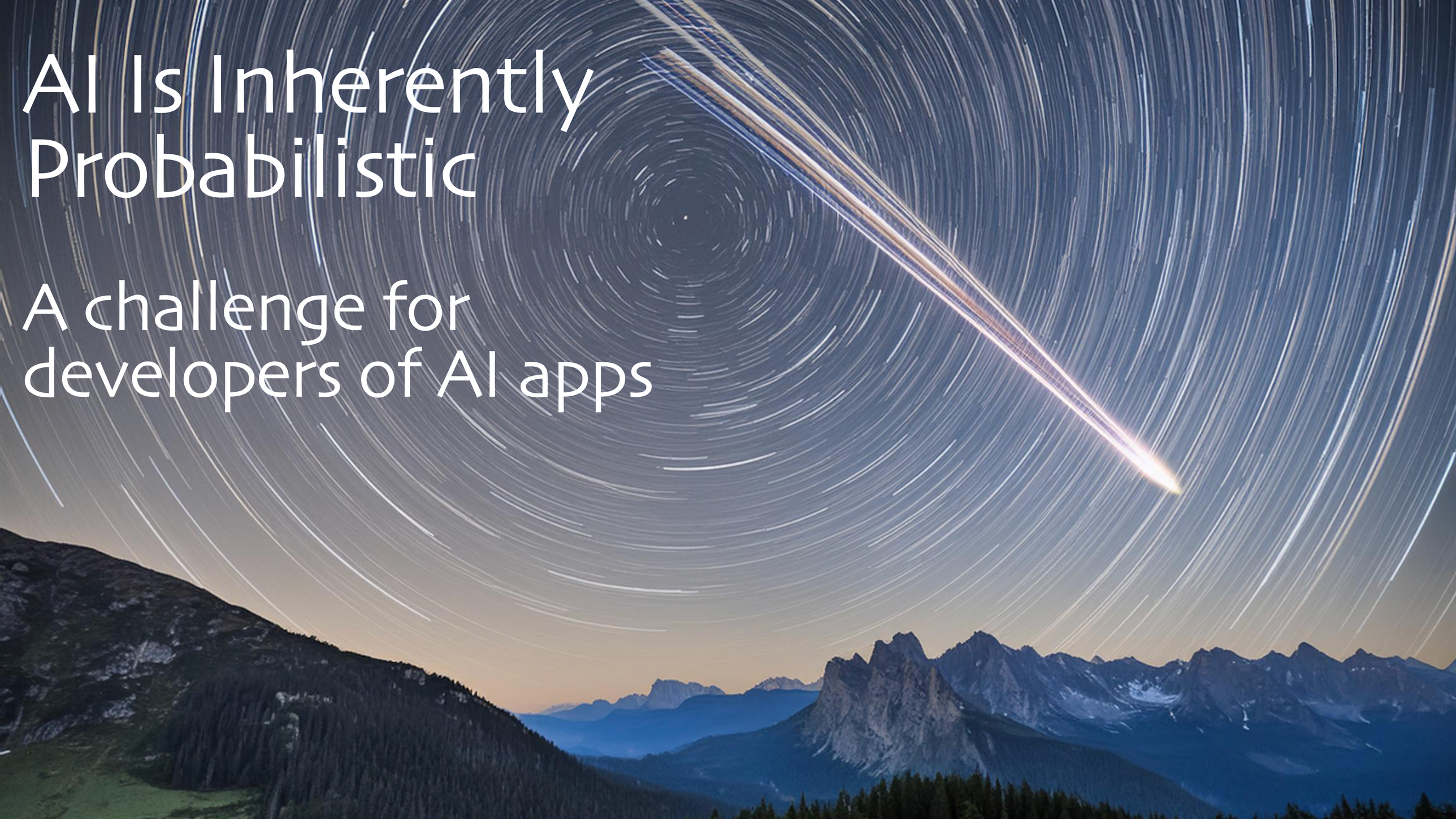
The Challenges to Success

The Challenges to Success

- AI Is Inherently Probabilistic
- Alignment
- Regulations and Policy
- Total Cost of Ownership

AI Is Inherently Probabilistic

A challenge for developers of AI apps



Developers expect software to be deterministic[‡]:

- The same input → the same output.
  - e.g., $\sin(\pi) = -1$
- The output is different? Something is broken!
- Developers rely on determinism to help ensure correctness and reproducibility.

[‡] Distributed systems break this clean picture.

Developers expect software to be deterministic‡:

- The s
  - e.g.
- The c oken!
- Deve ensure corre

Put another way, the determinism makes it easier to specify the system invariants, what should remain true from one iteration to the next.

‡ Distributed systems break this clean picture.

Generative models are probabilistic‡:

- The same prompt → **different** output.
  - chatgpt("Write a poem") → insanity
- How does a developer write a repeatable, reliable, test when she doesn't have determinism? Specifically,
  - Is that new model actually better or worse than the old model?
  - Did any regressions in behavior occur?

"Insanity is doing the same thing over and over again and expecting different results." — not Einstein

‡ A tunable "temperature" controls how probabilistic.

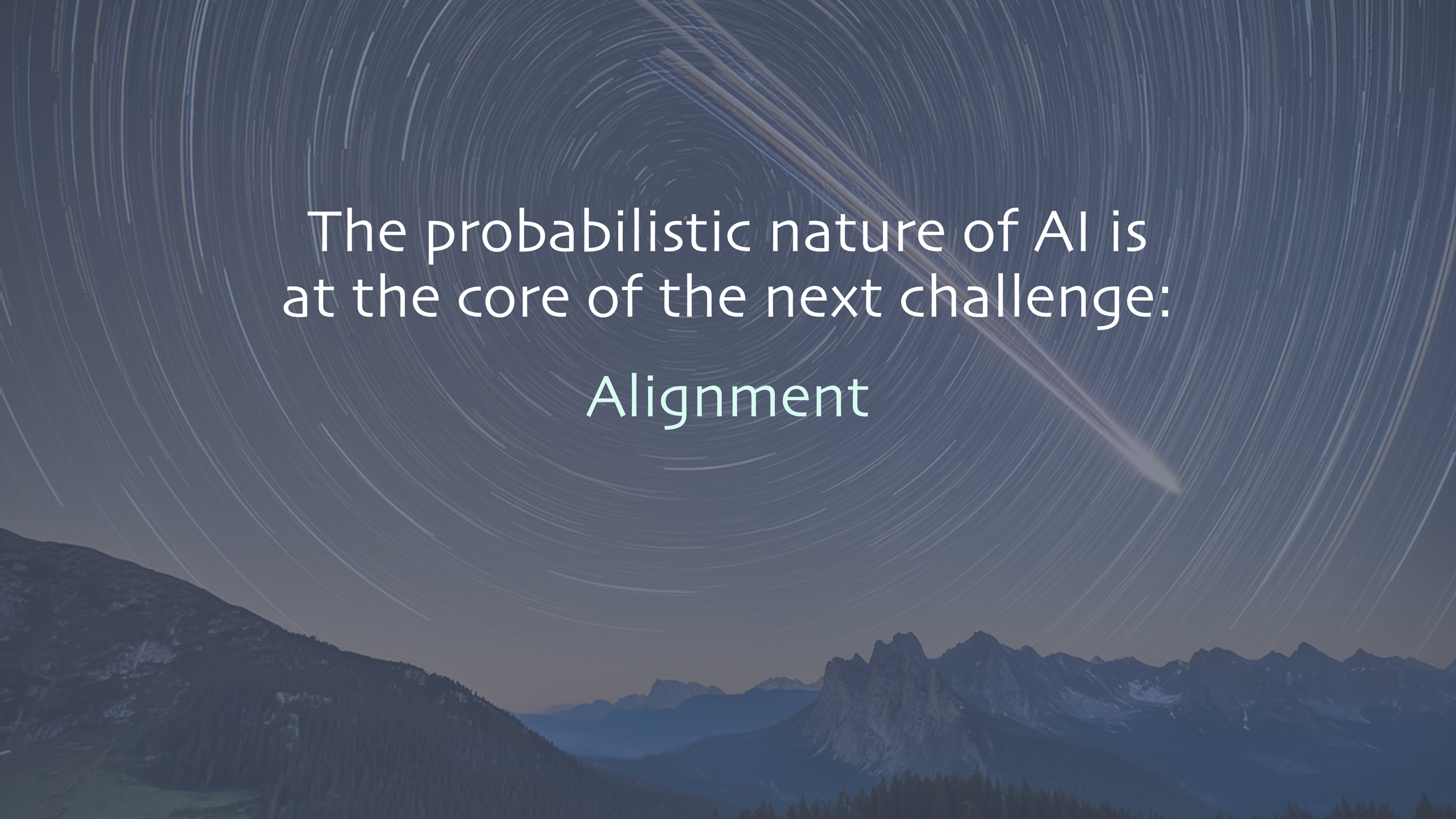
Generative models are probabilistic‡:

- The same prompt → **different** output.
  - cha
- How
  repe
  does
  - Is t
    wo
  - Did

Put another way, the system invariants, are not clear and therefore, much less enforceable.

"Insanity is doing the same thing over and over again and expecting different results." — not Einstein

‡ A tunable "temperature" controls how probabilistic.



The probabilistic nature of AI is at the core of the next challenge:

Alignment

Alignment

Alignment

Alignment - Assuring that the model or AI application works as intended, i.e., that the results satisfy requirements for:

- Usefulness for user goals
- Secure
- Free of bias
- Free of objectionable speech and concepts
- Free of copyrighted material
- Factually correct, i.e., free of hallucinations

Alignment

Alignment - Assurin
application works a
results satisfy requi

- Usefulness for u
- Free of bias
- Free of objectio
- Free of copyrigh
- Factually correct

Alignment is the hardest problem blocking broader adoption of Gen AI.

Hallucinations

Hallucinations remind us that context matters for alignment. What are your users' intentions and requirements?

Hallucinations

However, hallucinations **are** acceptable for:

- Tools for creative pursuits
  - Stories and scripts
  - Images and videos
- But copyright infringement is important.
- (I won't mention the concerns about impacting jobs for creatives…)

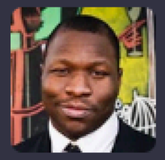
Hallucinations

But, hallucinations **are not** acceptable for:

- Customer service chatbots
- Recommenders, classifiers, etc. for Medical, legal, financial, …
- Search engines
- Resume writers
- Coding assistants

But these are the most hyped GenAI use cases!
What can we do?

What Actually Works?



Dare Obasanjo
@carnage4life@mas.to

There is a big difference between tech as augmentation versus automation. Augmentation (think Excel and accountants) benefits workers while automation (think traffic lights versus traffic wardens) benefits capital.

LLMs are controversial because the tech is best at augmentation but is being sold by lots of vendors as automation.

Jun 10, 2024, 10:31 · · Ivory for iOS · 109 · 188

https://mas.to/@carnage4life/112593042823322764

Emphasize Augmentation

- Keep humans in the loop, but improve productivity:
  - Distilling information more quickly
  - Translating between human and "domain languages"
    - SQL, Python, but also domain jargon (medical, finance, science, …)
  - Use complementary tools
    - Use deterministic tools for factual accuracy, logical reasoning, and planning
    - What Agent frameworks enable

Emphasize Augmentation

A.I. Needs Copper. It Just Helped to Find Millions of Tons of It.

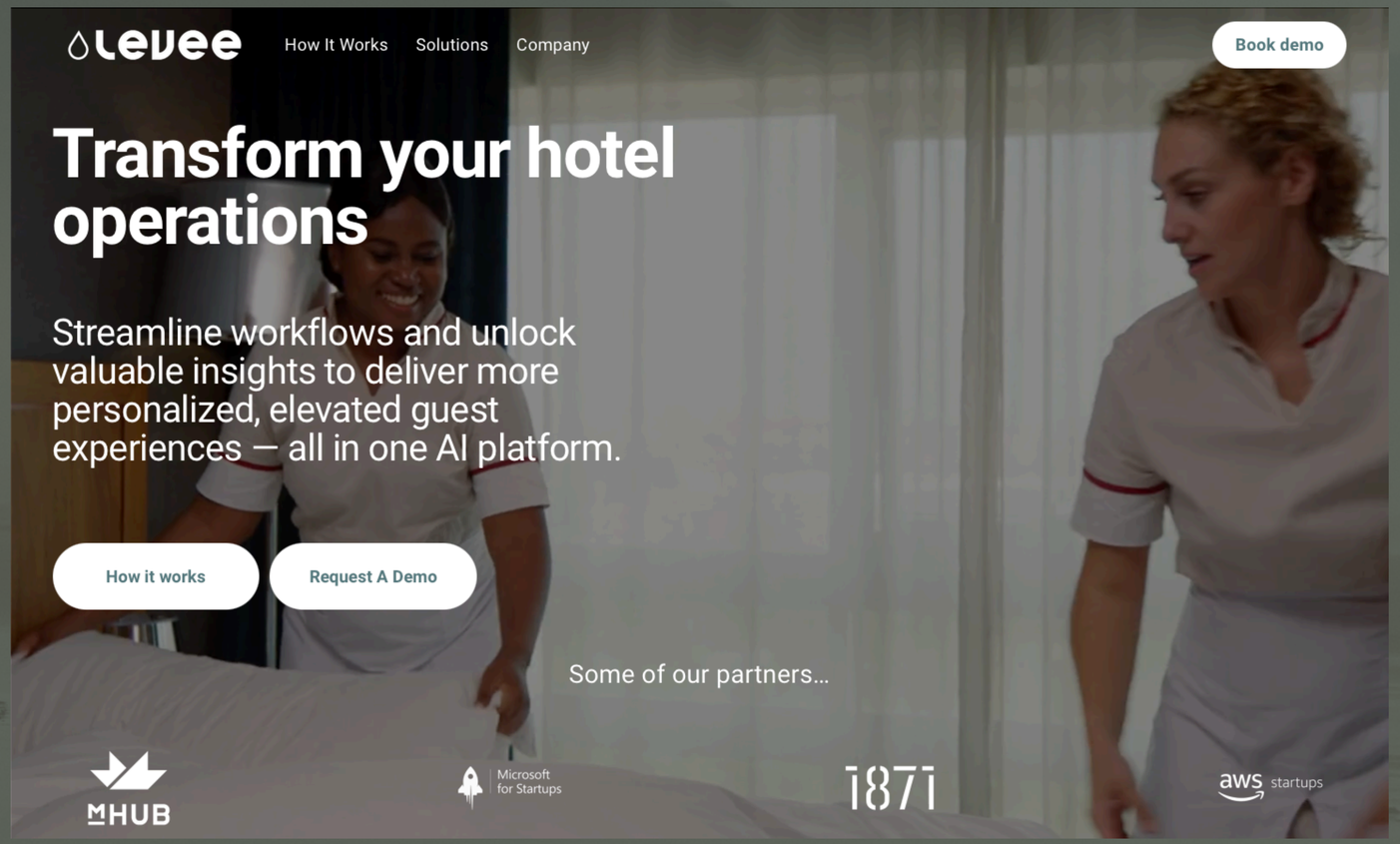
An exploration site run by KoBold Metals in Chililabombwe, Zambia, in June. Zinyange Auntor
York Times

The deposit, in Zambia, could make billions for Silicon Valley, provide minerals for the energy transition and help the United States in its rivalry with China.

Train model with geological and mining data to predict where the copper is likely to be

https://www.nytimes.com/2024/07/11/climate/kobold-zambia-copper-ai-mining.html

Emphasize Augmentation



https://www.levee.biz/

Levee uses machine vision to augment productivity for hotel housekeeping staff

Regulations and Policy

Safety Concerns

THE WHITE HOUSE

MENU

OCTOBER 30, 2023

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

BRIEFING ROOM › PRESIDENTIAL ACTIONS

By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:

Topics › Digital › Artificial intelligence › EU AI Act: first regulation on artificial intelligence

EU AI Act: first regulation on artificial intelligence

The use of artificial intelligence in the EU will be regulated by the AI Act, the world's first comprehensive AI law. Find out how it will protect you.

Published: 08-06-2023

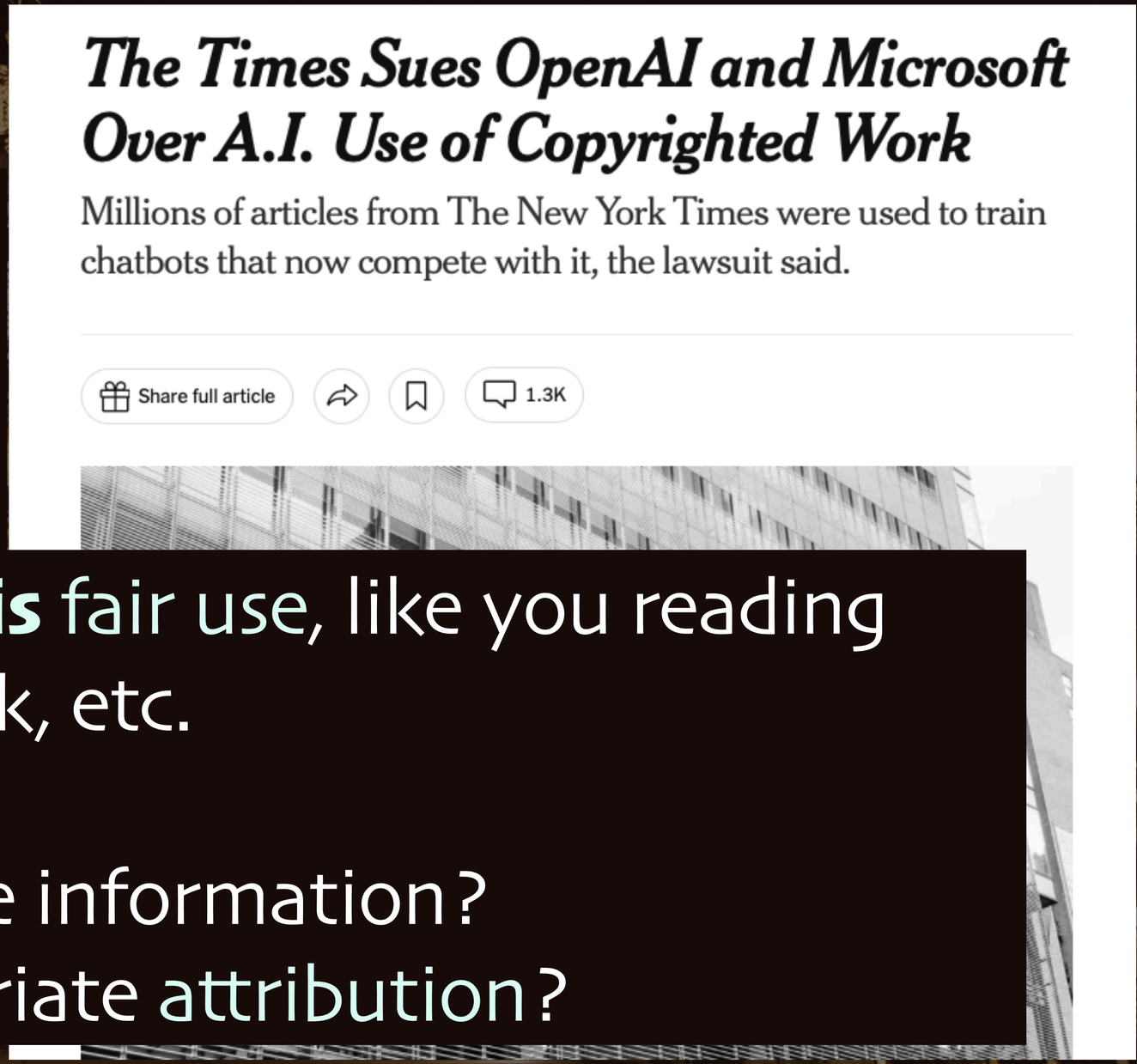
Last updated: 18-06-2024 - 16:29

6 min read

- whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-o artificial-intelligence/
- europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

Legal

Is it fair use to train with copyrighted data?



The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

Millions of articles from The New York Times were used to train chatbots that now compete with it, the lawsuit said.

Share full article 1.3K

- Some legal experts say, it **is** fair use, like you reading the NY Times, WSJ, a book, etc.
- What matters is:
  - How did you acquire the information?
  - Did you provide appropriate attribution?

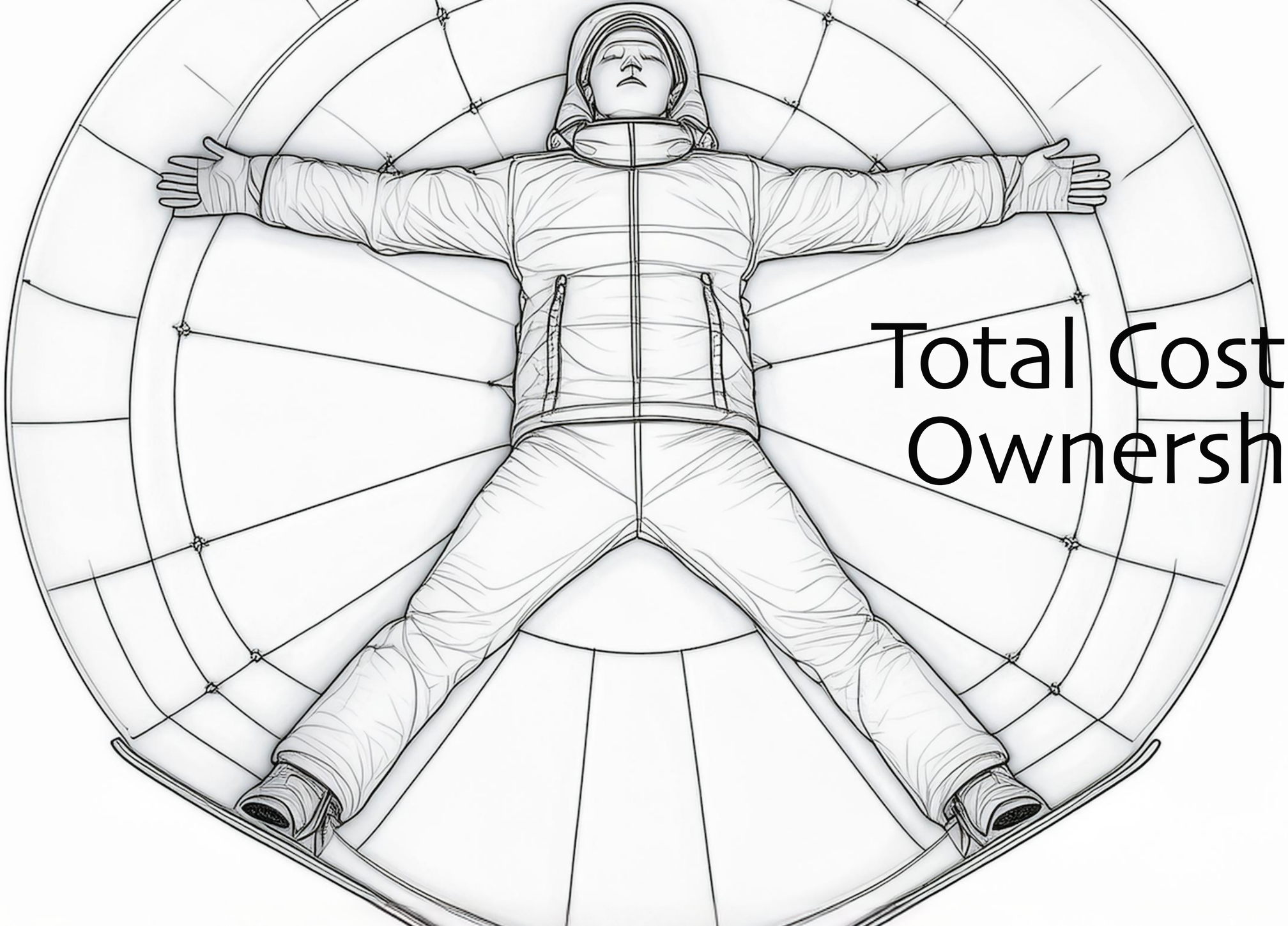
nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html

Question:

Can AI-generated content be copyrighted?

- "…, in the United States, copyright laws do not protect works created solely by a machine. But if an individual can demonstrate substantial human involvement in its creation, then it is plausible they may receive copyright protection."
- But if model training (prev. slide) is treated like a human activity, shouldn't creating content also be treated this way?

research.aimultiple.com/generative-ai-copyright/



Total Cost of Ownership

Generative AI Is Expensive

- TCO for Gen AI inference is expensive more than other services.

McKinsey: https://ceros.mckinsey.com/genai-cost-interactive-desktop/p/1

Estimated total cost of ownership for different archetypes

Taker	Shaper	Maker

Example use case

Customer service chatbot fine-tuned with sector-specific knowledge and chat history

Estimated total cost of ownership

~$2.0 million to $10.0 million, one-time unless model is fine-tuned further

- Data and model pipeline building: ~$0.5 million. Costs include 5 to 6 machine learning engineers and data engineers working for 16 to 20 weeks to collect and label data and perform data ETL.[1]
- Model fine-tuning[2]: ~$0.1 million to $6.0 million per training run[3]
  - Lower end: costs include compute and 2 data scientists working for 2 months
  - Upper end: compute based on public closed-source model fine-tuning cost
- Plug-in-layer building: ~$1.0 million to $3.0 million. Costs include a team of 6 to 8 working for 6 to 12 months.

~$0.5 million to $1.0 million, recurring annually

- Model inference: up to ~$0.5 million recurring annually. Assume 1,000 chats daily with both audio and texts.
- Model maintenance: ~$0.5 million. Assume $100,000 to $250,000 annually for MLOps

Forbes

FORBES > INNOVATION > AI

Generative AI Breaks The Data Center: Data Center Infrastructure And Operating Costs Projected To Increase To Over $76 Billion By 2028

Jim McGregor Contributor
Tirias Research Contributor Group

Follow

2

May 12, 2023, 04:33pm EDT

Forbes: link

Harvard Business Review - What CEOs Need to Know About the Costs of Adopting GenAI:
https://hbr.org/2023/11/what-ceos-need-to-know-about-the-costs-of-adopting-genai

One Solution: Smaller Models

In 2023 we learned useful model size tradeoffs:

- Big models:
  - ✓ More generalizable
  - ✓ Highest benchmark scores
  - ✗ Much higher costs
  - ✗ High carbon footprint
- Small models:
  - ✗ Less generalizable
  - ✓ Easy to tune to be very good for specific applications
  - ✓ Much lower costs
  - ✓ Lower carbon footprint

One Solution: Smaller Models

- Mixture of Experts
  - Combine several smaller, cheaper models match performance of one large model

Few organizations train models from scratch. Instead, they pick a good, "open" model and tune it for their needs.

Generative AI in Five Years?

A current perspective:
https://www.nytimes.com/2024/07/24/opinion/ai-annoying-future.html

What About Chatbots?

Will Chatbots rule or are they a temporary "fad"?

- ChatGPT and other general-purpose, heavily-engineered chatbots are already great for many human tasks, like creative work, simple coding needs, etc.
- Enterprise chatbots are mostly terrible now, but…
- Voice response systems predate LLMs:
  - They should get better with LLMs + "smart" application patterns.

What Problems Are Already Being Solved?

Hardware costs and energy demands will drop:

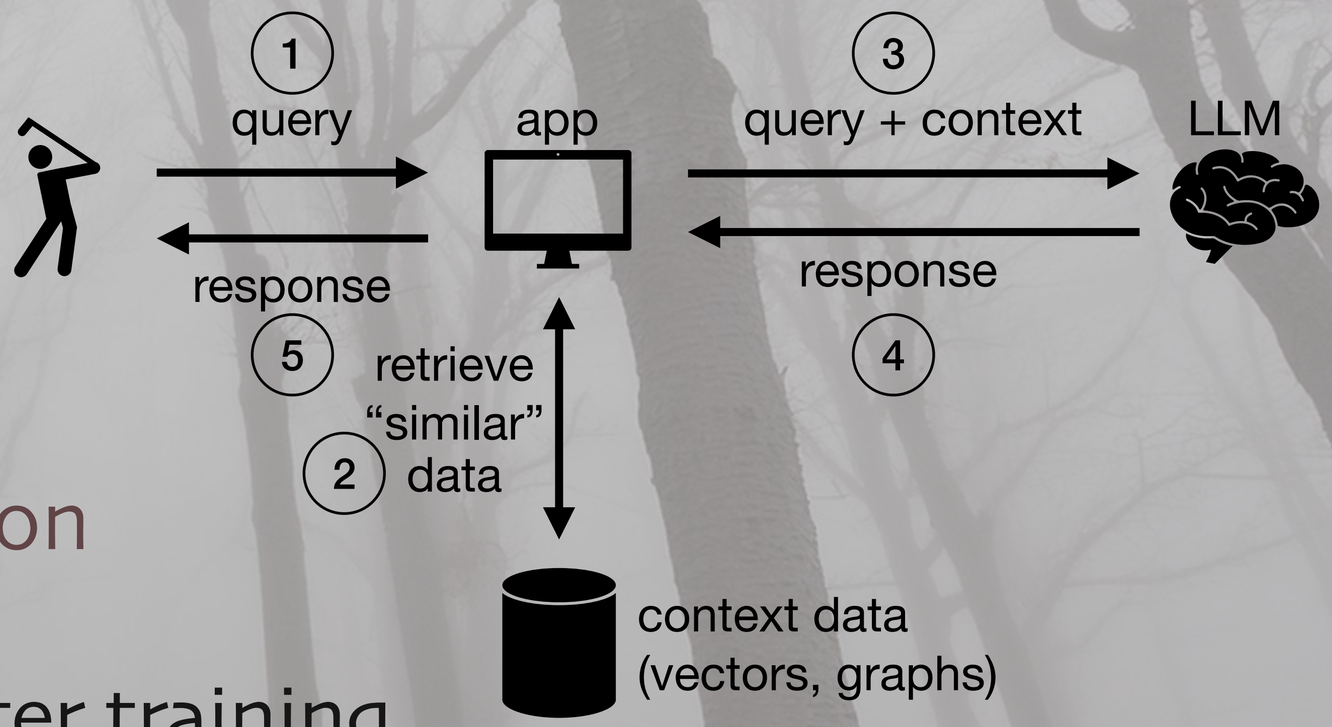
- New, more efficient accelerator architectures
  - GPU alternatives from AMD, Intel, Cerebras, Google, Microsoft, AWS, Apple, IBM, …
- Alternative model architectures to Transformers?
  - See this Reddit post
- Optimizations for efficient training, tuning, and inference

What Problems Are Already Being Solved?

Application architectures will not rely solely on models:

- General-purpose, generative models will always hallucinate.
- We are combining models with other techniques and tools…
- … let's look at the current state of the art.

Retrieval-Augmented Generation (RAG)



First generation tool integration

- Improves alignment
- Incorporates knowledge after training
- Incorporates proprietary knowledge

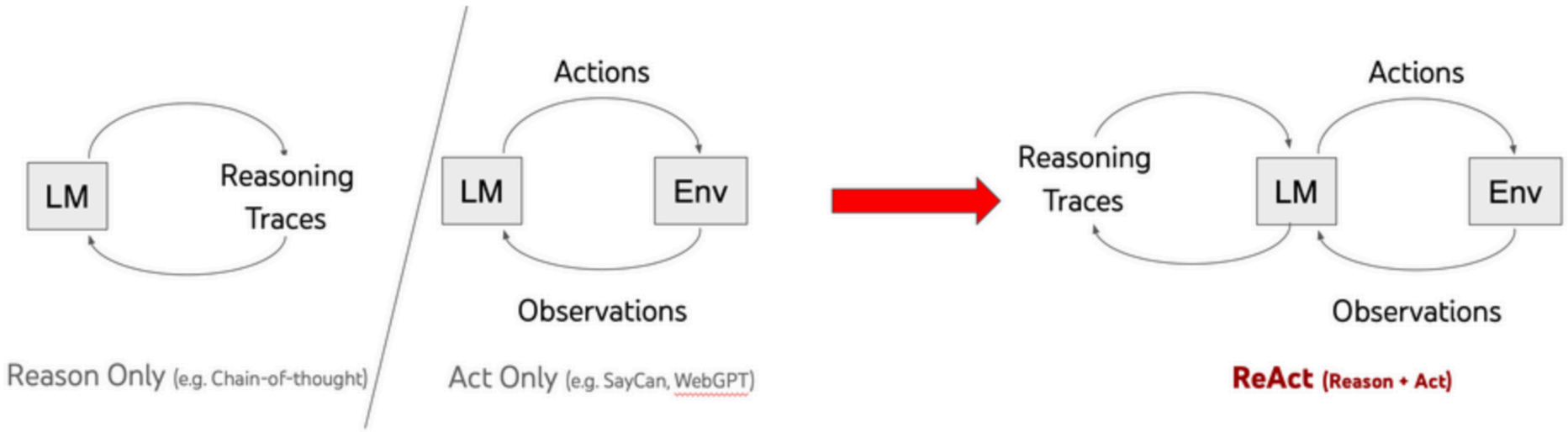
Agents Example: ReAct

https://react-lm.github.io

ReAct: Synergizing Reasoning and Acting in Language Models

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, Yuan Cao

[Paper] [Code] [Blogpost] [BibTex]

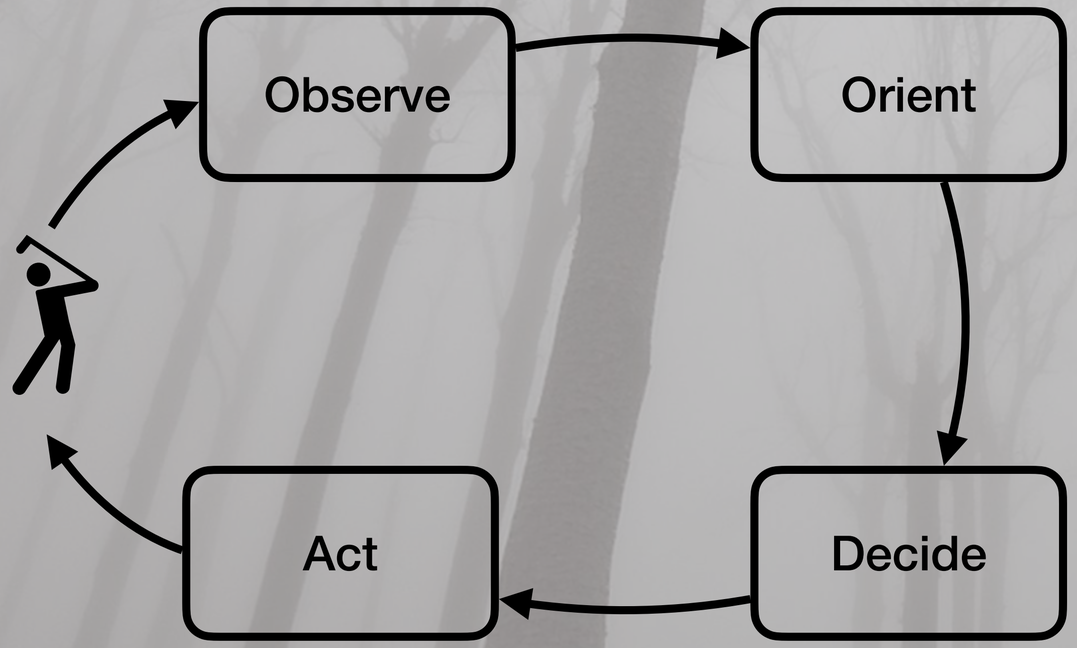


Language models are getting better at reasoning (e.g. chain-of-thought prompting) and acting (e.g. WebGPT, SayCan, ACT-1), but these two directions have remained separate.

ReAct asks, what if these two fundamental capabilities are combined?

https://react-lm.github.io/

Agents Example: OODA/OpenSSA



https://www.openssa.org/
https://thealliance.ai/blog/advancing-domain-specific-qa-the-ai-alliances-guid

RAG, ReAct, OODA, …

These are today's state of the art. We will have more sophisticated approaches in ~five years.

What Will Life Be Like?

The Matrix? Or will AI be a normal, ubiquitous part of daily life, like the Internet is today?

- Enhanced productivity in work and life
- … but with lingering concerns about safety, jobs, …

A revival of human writing, painting, photography, …

- We'll be sick of AI-generated content

Thank You!

- Visit thealliance.ai
- Let me know what you think!
  - dwampler@thealliance.ai
  - Mastodon and Bluesky: @deanwampler
  - Other talks: deanwampler.com/talks: use this ➡

deanwampler.com/talks

Extra Slides

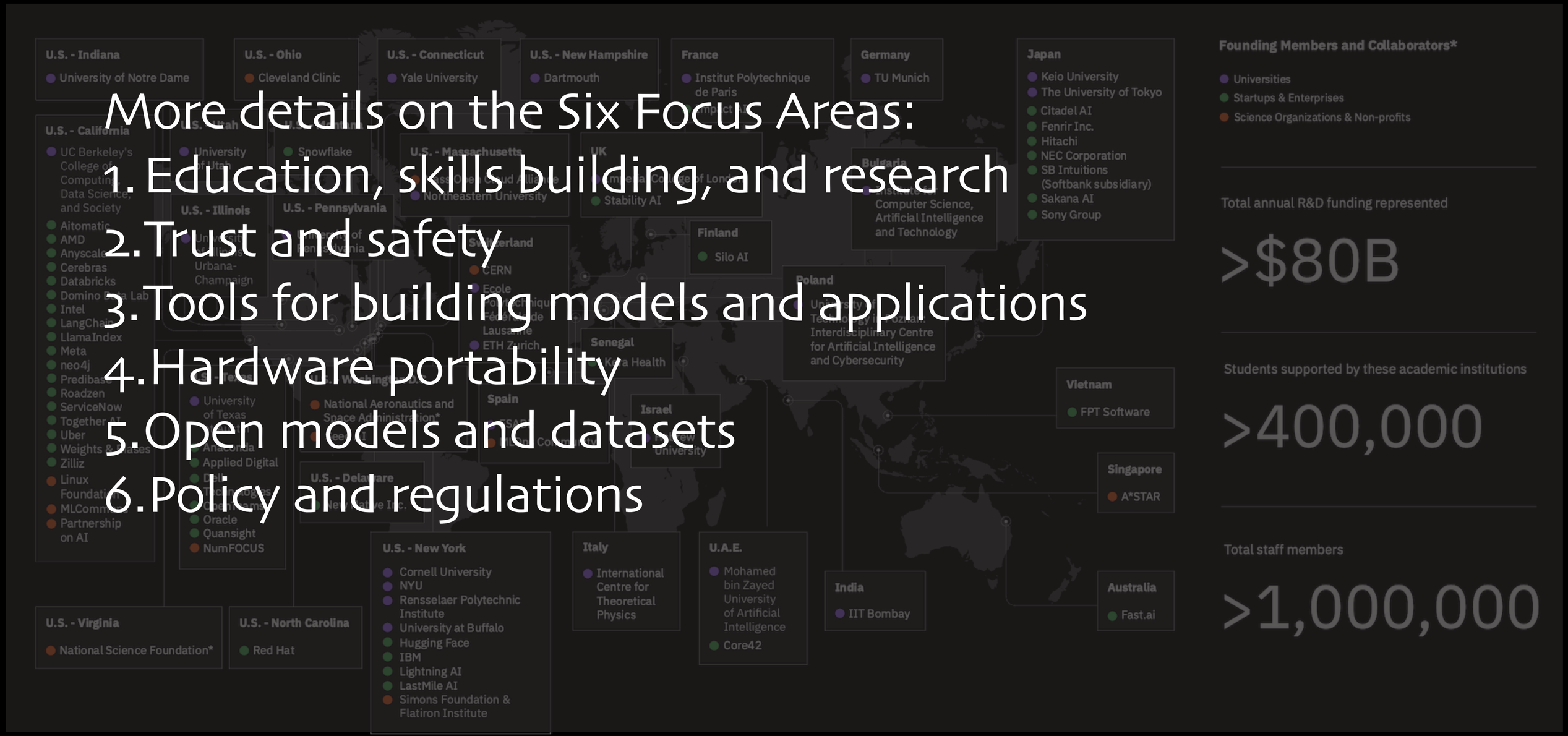
Notes

© Text 2023-2024, Dean Wampler, © Images 2004-2024, Dean Wampler, except where noted. Most of the images are based on my photographs (flickr.com/photos/deanwampler/), but they are manipulated by AI in some way. Where noted, the image was generated by Adobe Firefly with one of my pictures as a "reference image" for the style. For other images, I used Firefly to add elements to my photograph.

1. Title slide uses this Chicago Park image enhanced with Firefly: flickr.com/photos/deanwampler/53419199087/in/dateposted-public/
2. "Should AI be open or closed?" and the end "thank you" slide images were both generated by Firefly using the same sunset image taken from Clingmans Dome, Great Smoky Mountains NP as a reference image: flickr.com/photos/deanwampler/51664228468/in/album-72157720120215384/
3. "The Challenges to Success" image was generated by Firefly using this Tower of London image as a reference image: https://www.flickr.com/photos/deanwampler/30651106445/in/album-72157649394354046/
4. "AI Is Inherently Probabilistic", image generated by Firefly using this Wind River Range astro image as a reference image: flickr.com/photos/deanwampler/53004539434/in/album-72177720302185576/
5. "Alignment" image is an Oregon coast image enhanced with Firefly: flickr.com/photos/deanwampler/4850305877/in/album-72157624506732831/
6. "Regulation and Policy" image is a fake European government building where I used a night-time image of the Brussels City Hall as the reference image (not on Flickr).
7. "Total Cost of Ownership" was generated by Firefly where I asked for "Leonardo da Vinci's 'Vitruvian Man' as a snow angel." Having him in skiing gear was part of the output, not my prompt.
8. "Generative AI in Five Years?" Image was generated by Firefly using the same title slide Chicago Park image as a reference image, where I also requested the addition of a bigfoot.

Meet the AI Alliance

thealliance.ai

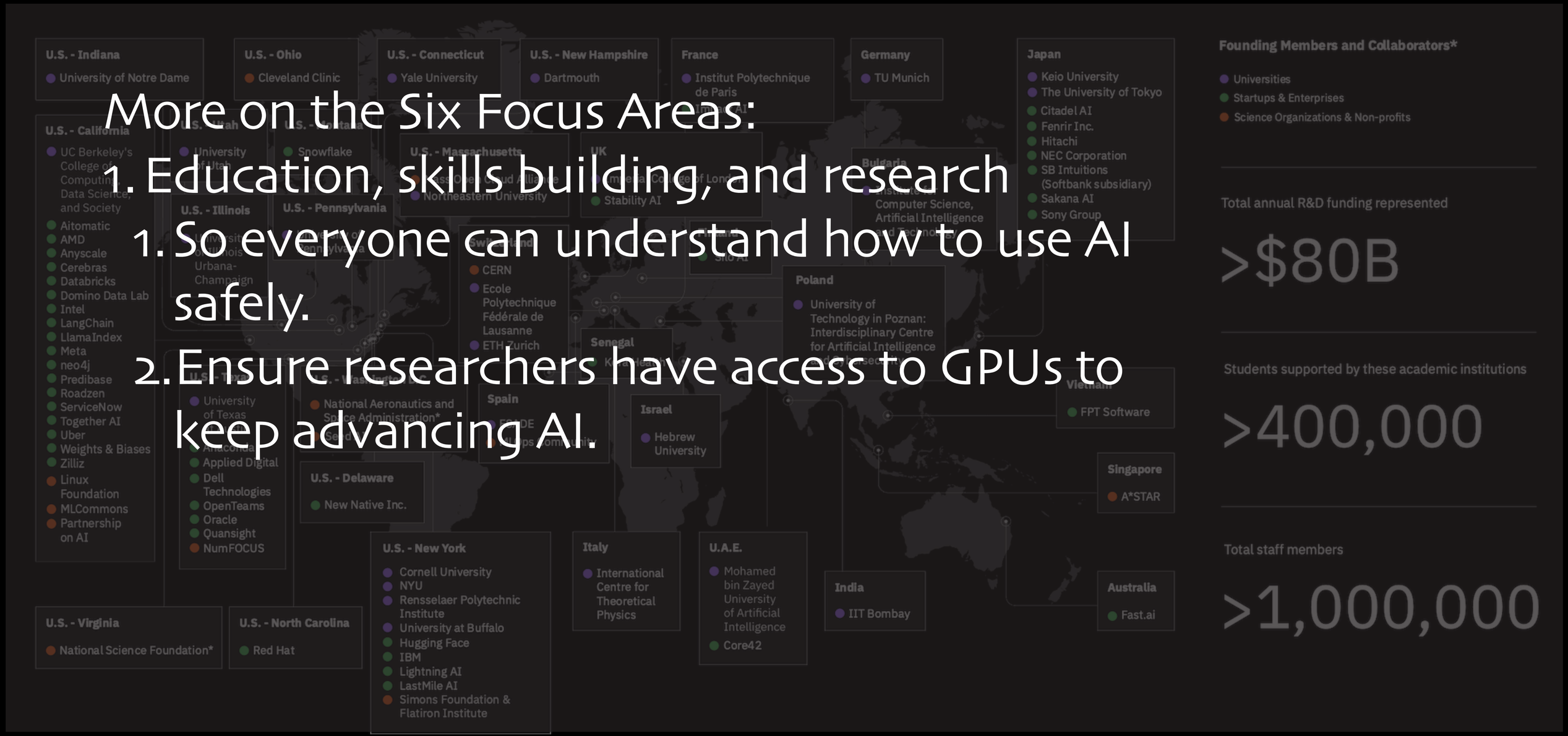


More details on the Six Focus Areas:

1. Education, skills building, and research
2. Trust and safety
3. Tools for building models and applications
4. Hardware portability
5. Open models and datasets
6. Policy and regulations

Meet the AI Alliance

thealliance.ai

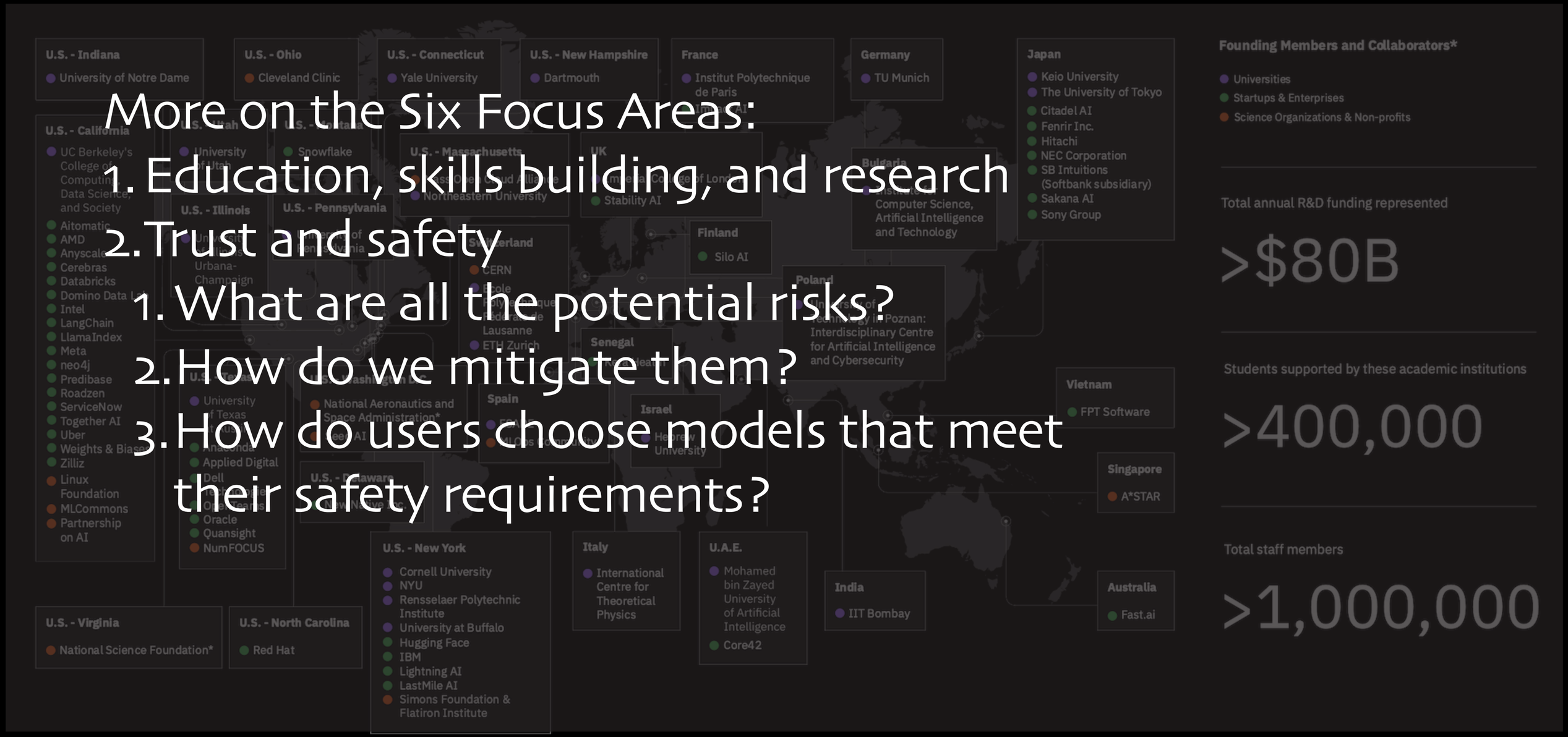


More on the Six Focus Areas:

1. Education, skills building, and research
   1. So everyone can understand how to use AI safely.
   2. Ensure researchers have access to GPUs to keep advancing AI.

Meet the AI Alliance

thealliance.ai

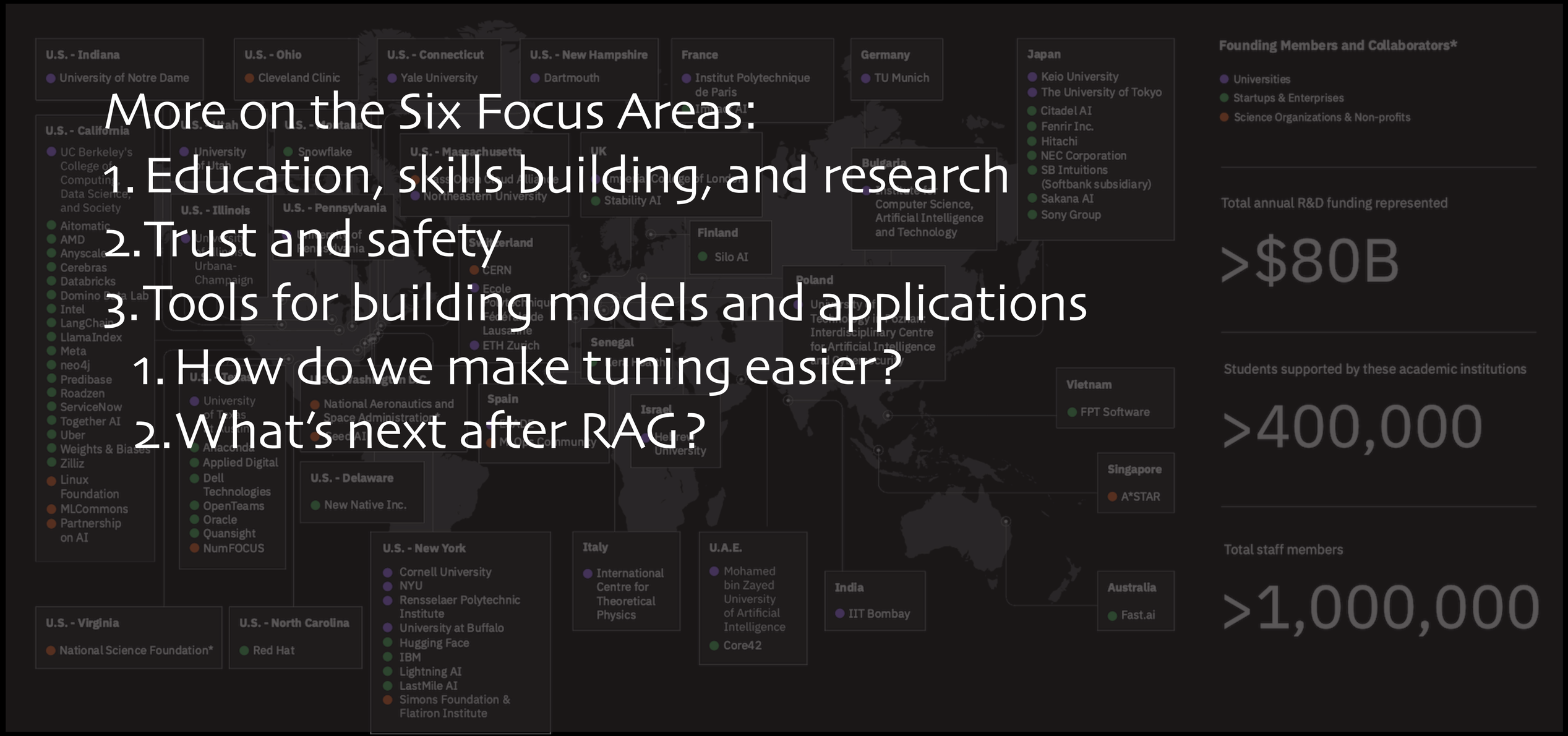


More on the Six Focus Areas:

1. Education, skills building, and research
2. Trust and safety
   1. What are all the potential risks?
   2. How do we mitigate them?
   3. How do users choose models that meet their safety requirements?

Meet the AI Alliance

thealliance.ai

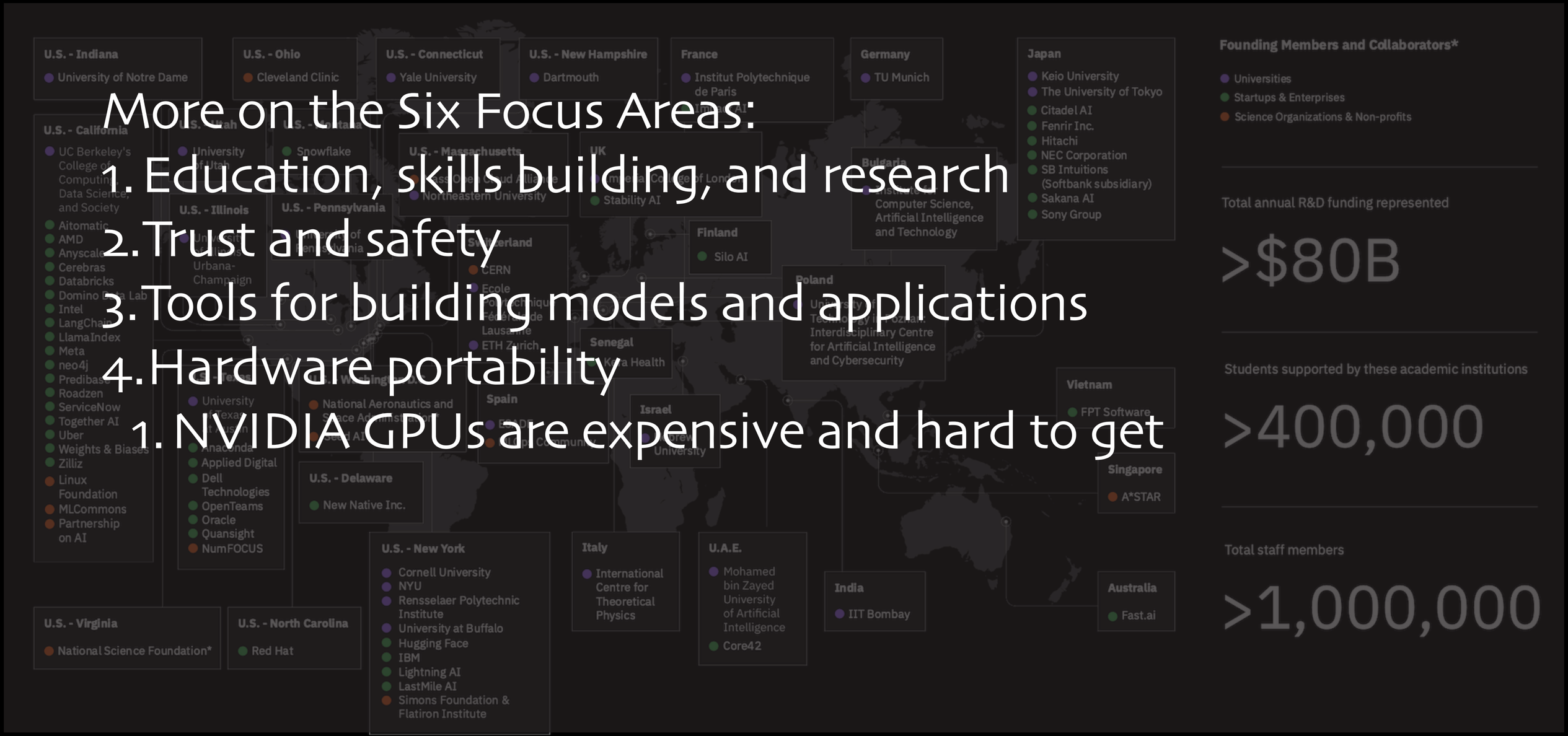


More on the Six Focus Areas:

1. Education, skills building, and research
2. Trust and safety
3. Tools for building models and applications
   1. How do we make tuning easier?
   2. What's next after RAG?

Meet the AI Alliance

thealliance.ai

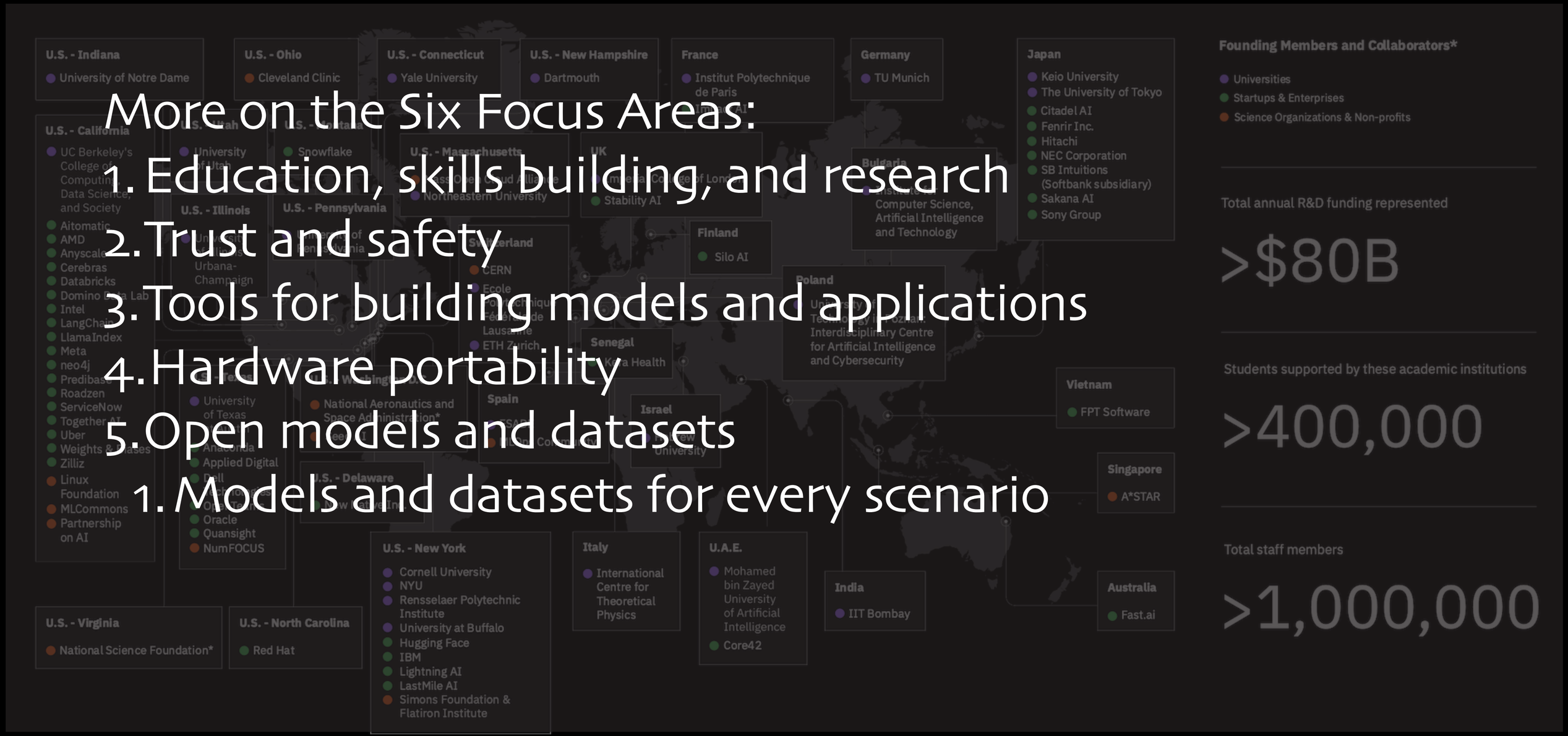


More on the Six Focus Areas:

1. Education, skills building, and research
2. Trust and safety
3. Tools for building models and applications
4. Hardware portability
   1. NVIDIA GPUs are expensive and hard to get

Meet the AI Alliance

thealliance.ai

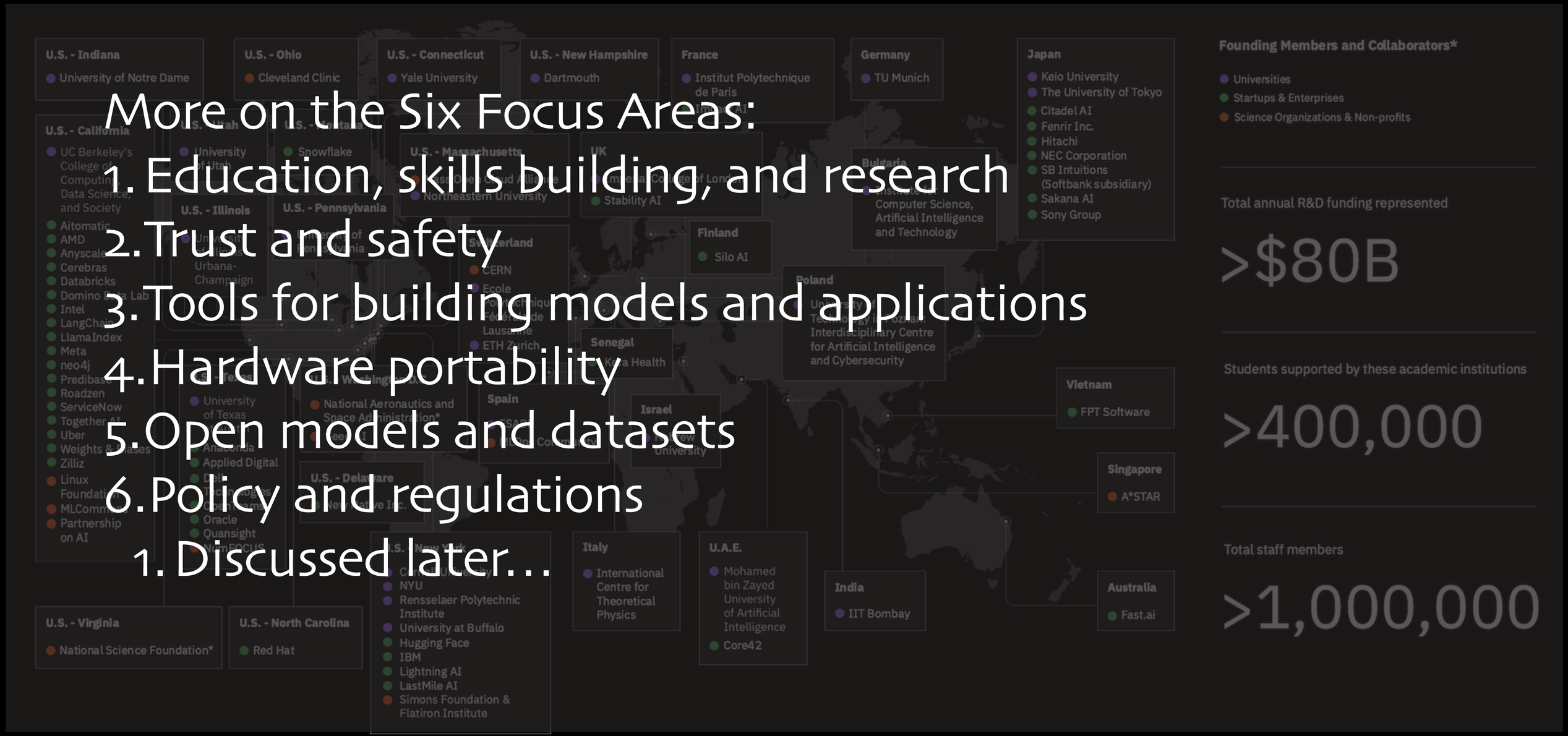


More on the Six Focus Areas:

1. Education, skills building, and research
2. Trust and safety
3. Tools for building models and applications
4. Hardware portability
5. Open models and datasets
   1. Models and datasets for every scenario

Meet the AI Alliance

thealliance.ai



More on the Six Focus Areas:

1. Education, skills building, and research
2. Trust and safety
3. Tools for building models and applications
4. Hardware portability
5. Open models and datasets
6. Policy and regulations
   1. Discussed later…