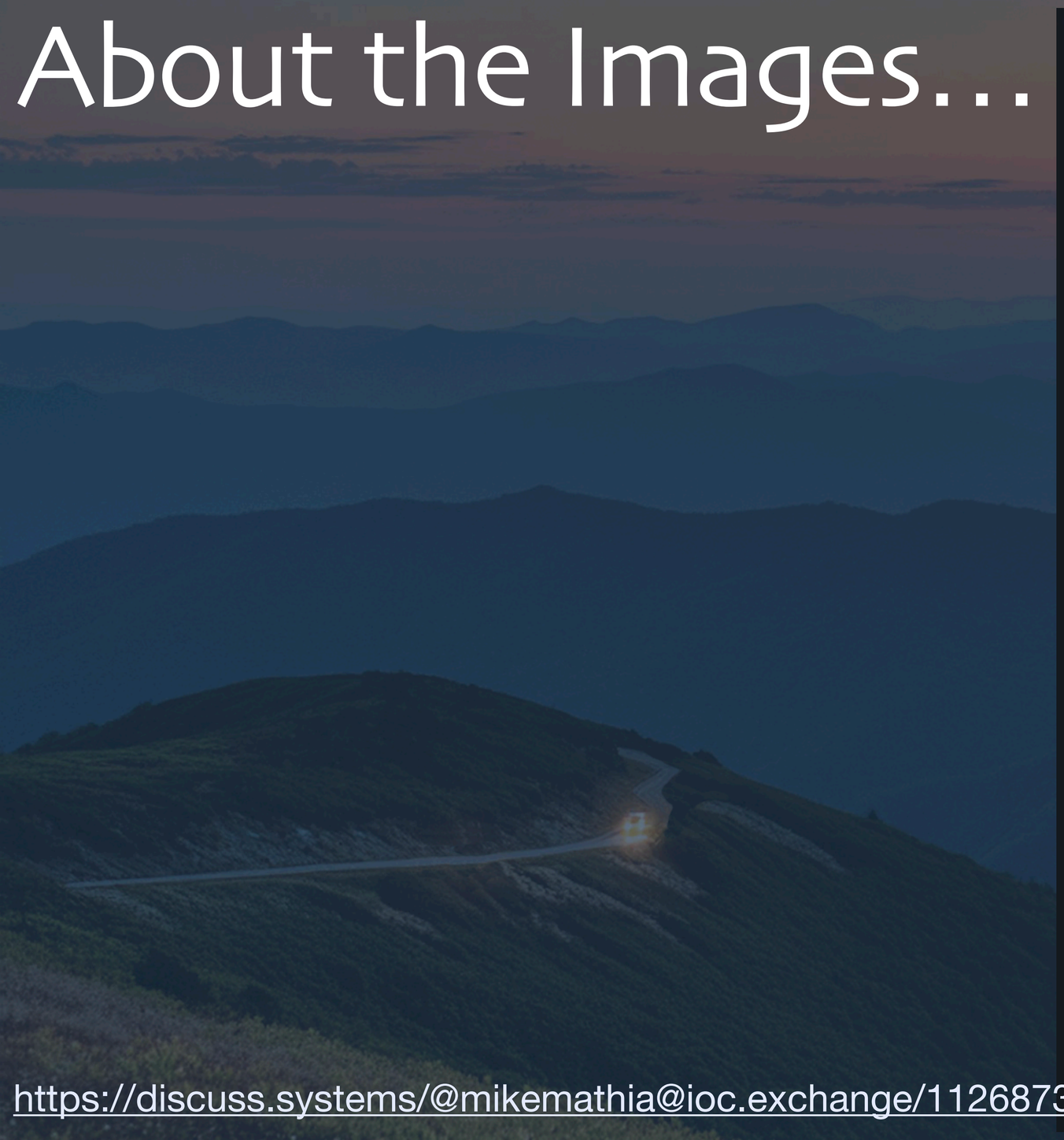
What Issues Are Blocking AI Adoption?

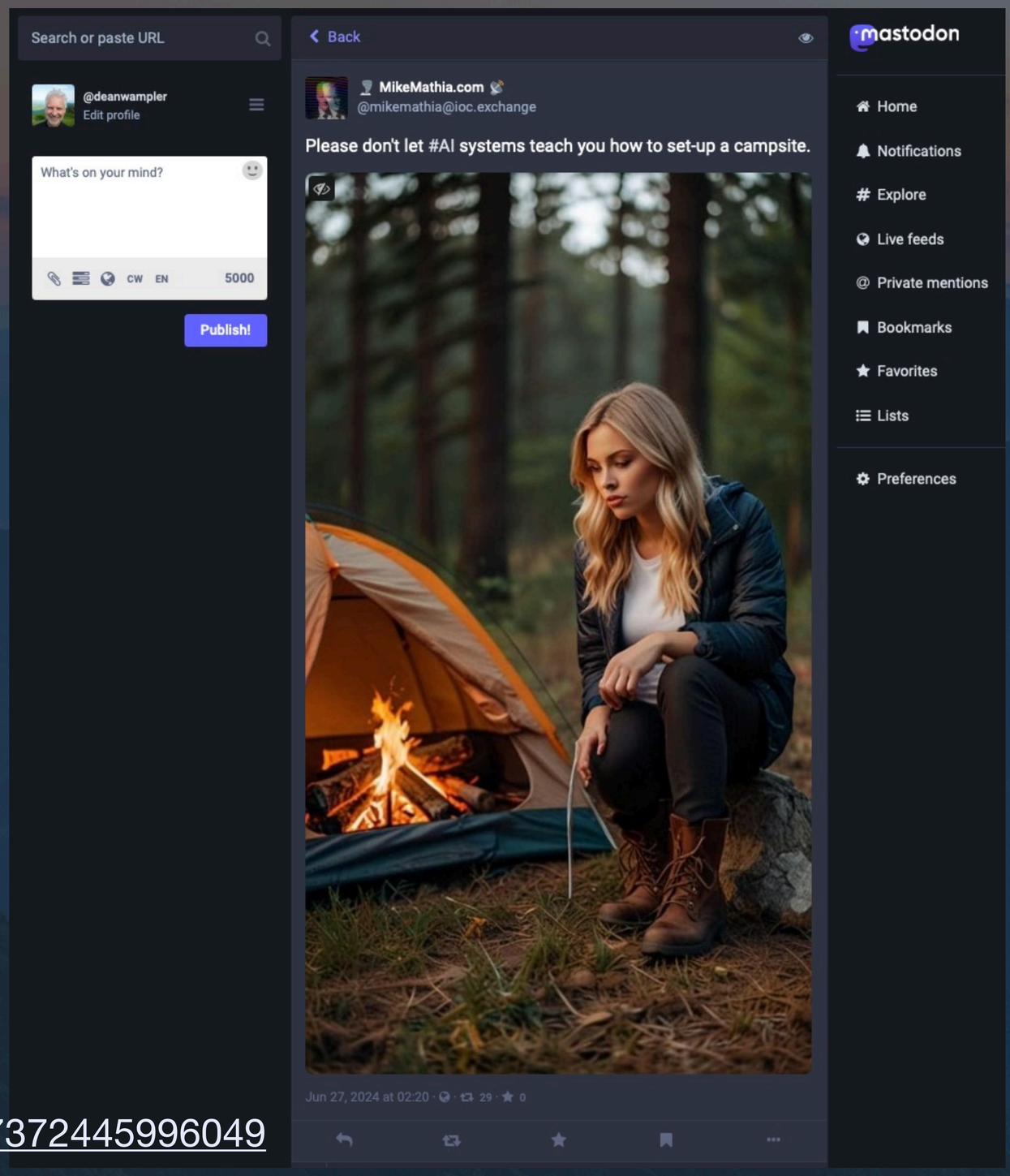
1871 AI Innovation Summit
June 27, 2024

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

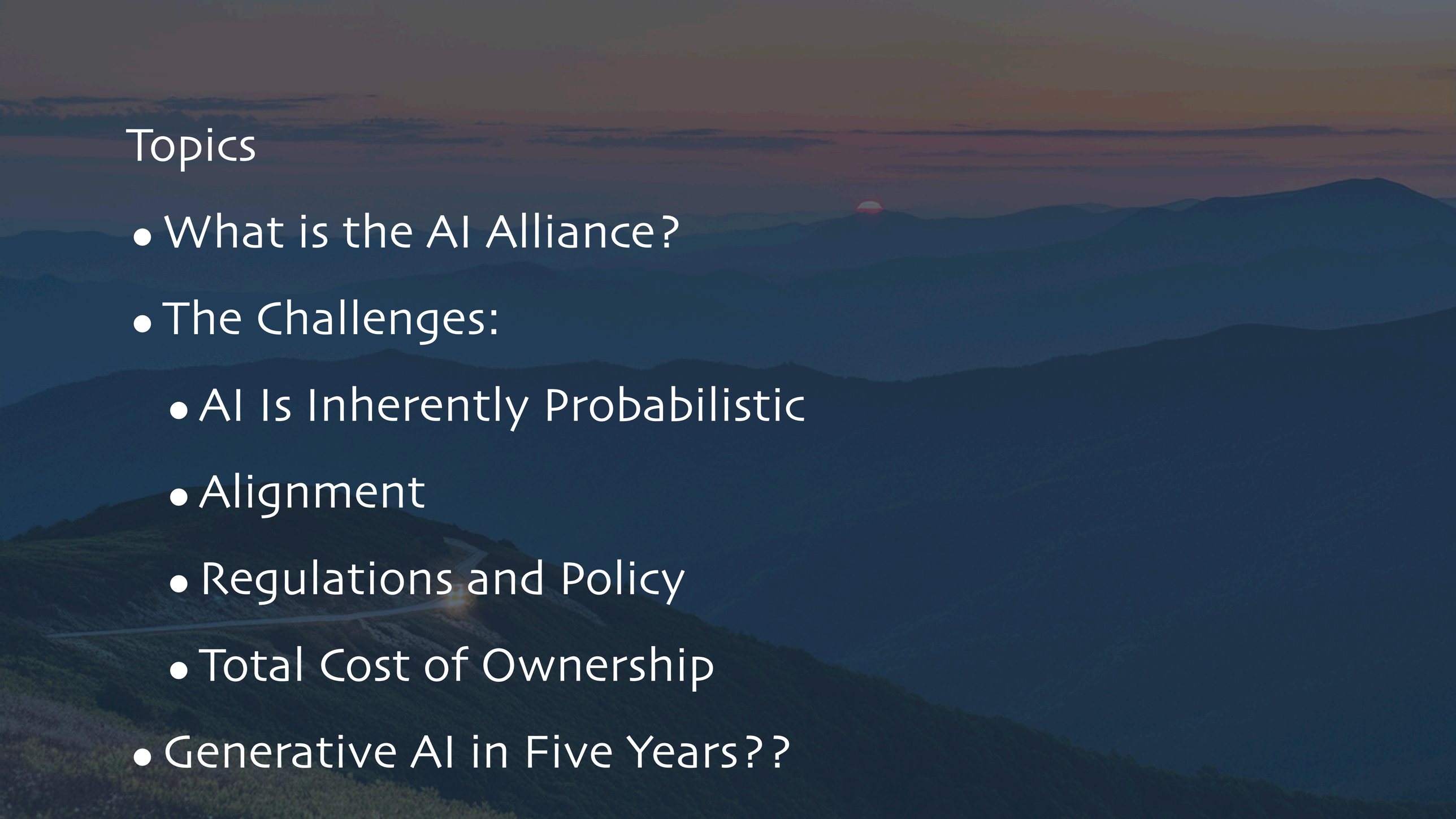
© 2004-2024, Dean Wampler, except where noted.

About the Images…





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



Topics

- What is the AI Alliance?
- The Challenges:
  - AI Is Inherently Probabilistic
  - Alignment
  - Regulations and Policy
  - Total Cost of Ownership
- 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.

Learn more ↓

The **AI Alliance**

thealliance.ai

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

Founding Members and Collaborators*

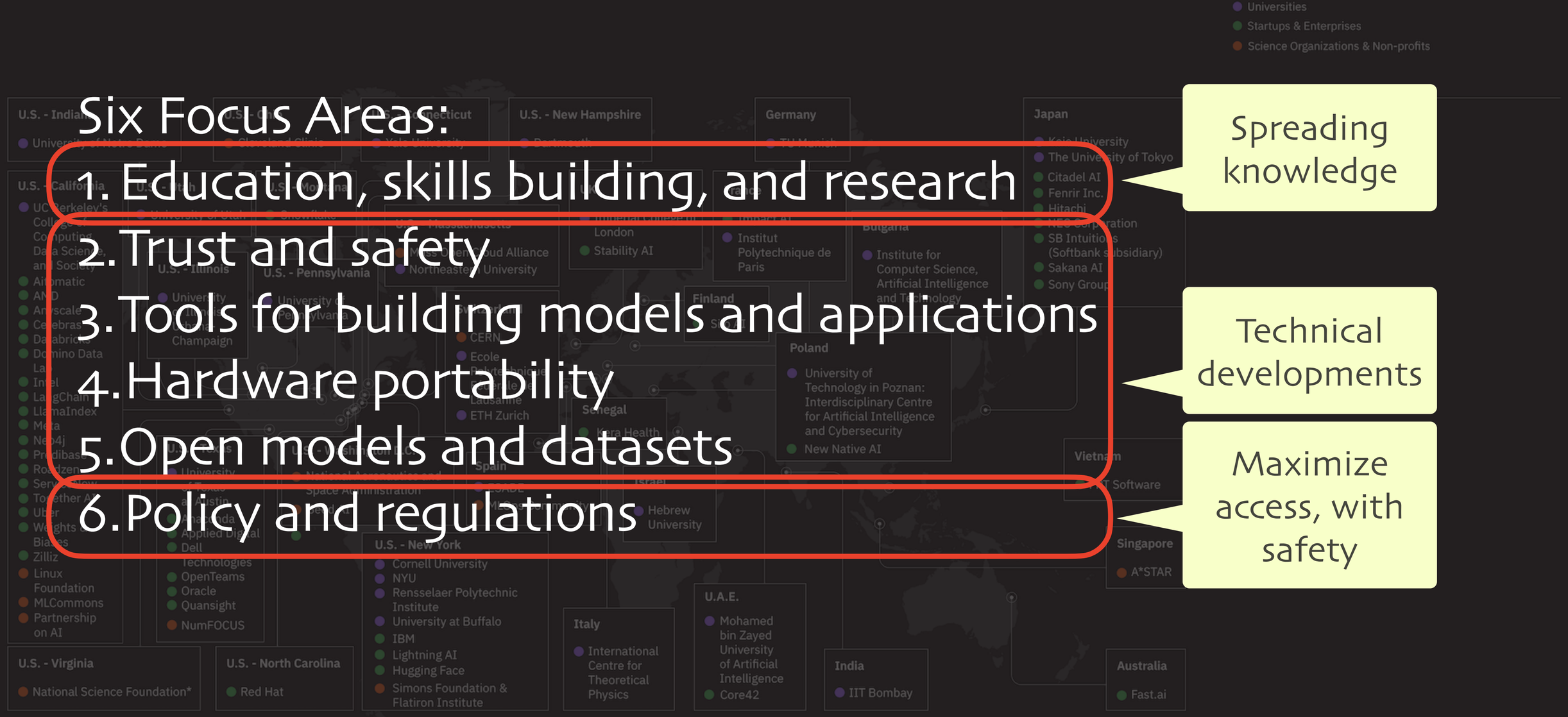
- Universities
- Startups & Enterprises
- Science Organizations & Non-profits

Diagram as of February.
>100 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

Technical developments

Maximize access, with safety

The Challenges

AI Is Inherently Probabilistic

Developers are accustomed to software systems that are deterministic (more or less‡).

- Calling a function with the same input returns the same output.
  - e.g., $\sin(\pi) = -1$
- When a result is different, then something has broken (a regression)!
- We have relied on determinism to help ensure quality and correctness.

‡ Distributed systems break this clean picture.

Generative models are inherently probabilistic (more or less‡).

- Calling a model with the same prompt returns different results.
  - e.g., `chatgpt("Write a poem")` -> insanity
- How do you write a repeatable test?
- Is that new model actually better or worse than the previous model?

‡ A tunable "temperature" controls how probabilistic.

Hence Generative AI developers have a previously unseen‡ challenge of how to ensure code quality and correctness.

The probabilistic nature is at the core of the next challenge: Alignment.

‡ Well, except for distributed systems again.

Alignment

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

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

Alignment - Assuring tha
application works as inte
results satisfy requiremer

- Usefulness for user go
- Free of bias
- Free of objectionable s
- Free of copyrighted m
- Factually correct, i.e., f

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

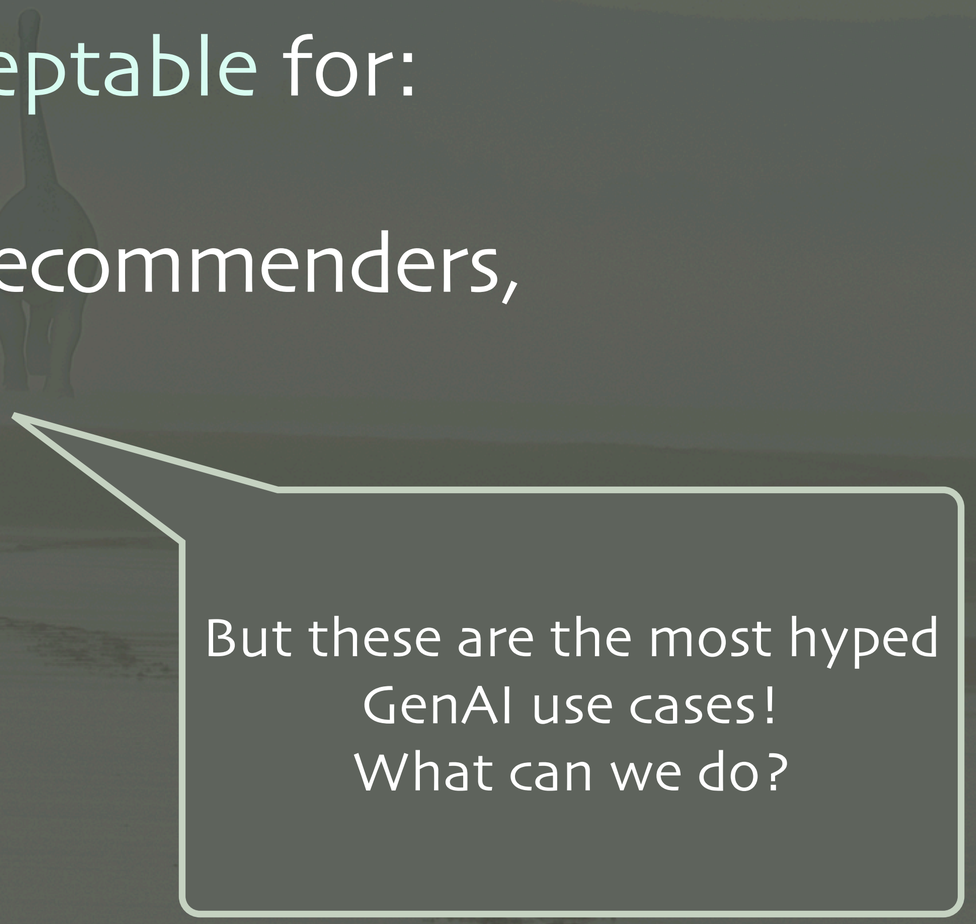
However, hallucinations **are** acceptable for:

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

Hallucinations

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

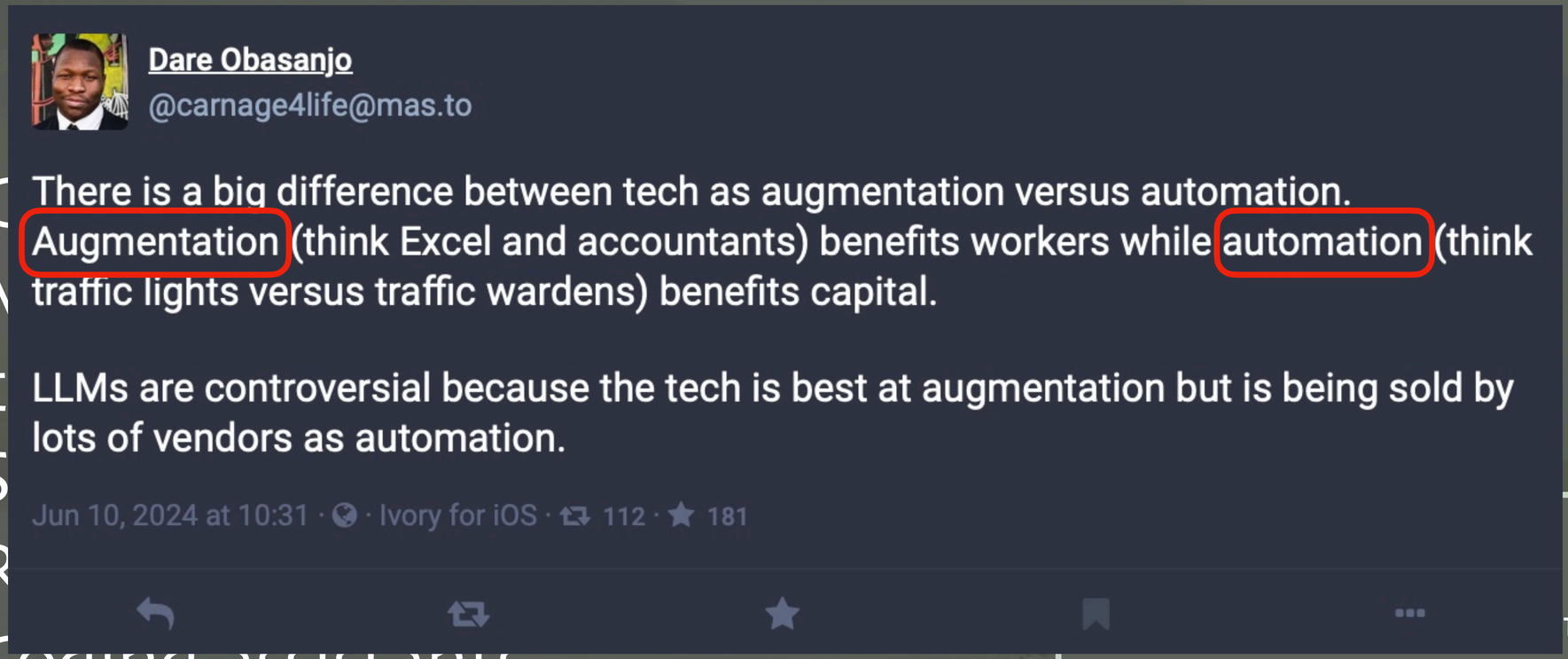
- Customer service chatbots
- Medical, legal, financial, … recommenders, classifiers, etc.
- Search engines
- Resume writers
- Coding assistants



Hallucinations

But,

- C
- A
  c
- S
- R
- Coding assistants



t hyped
GenAI use cases!
What can we do?

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

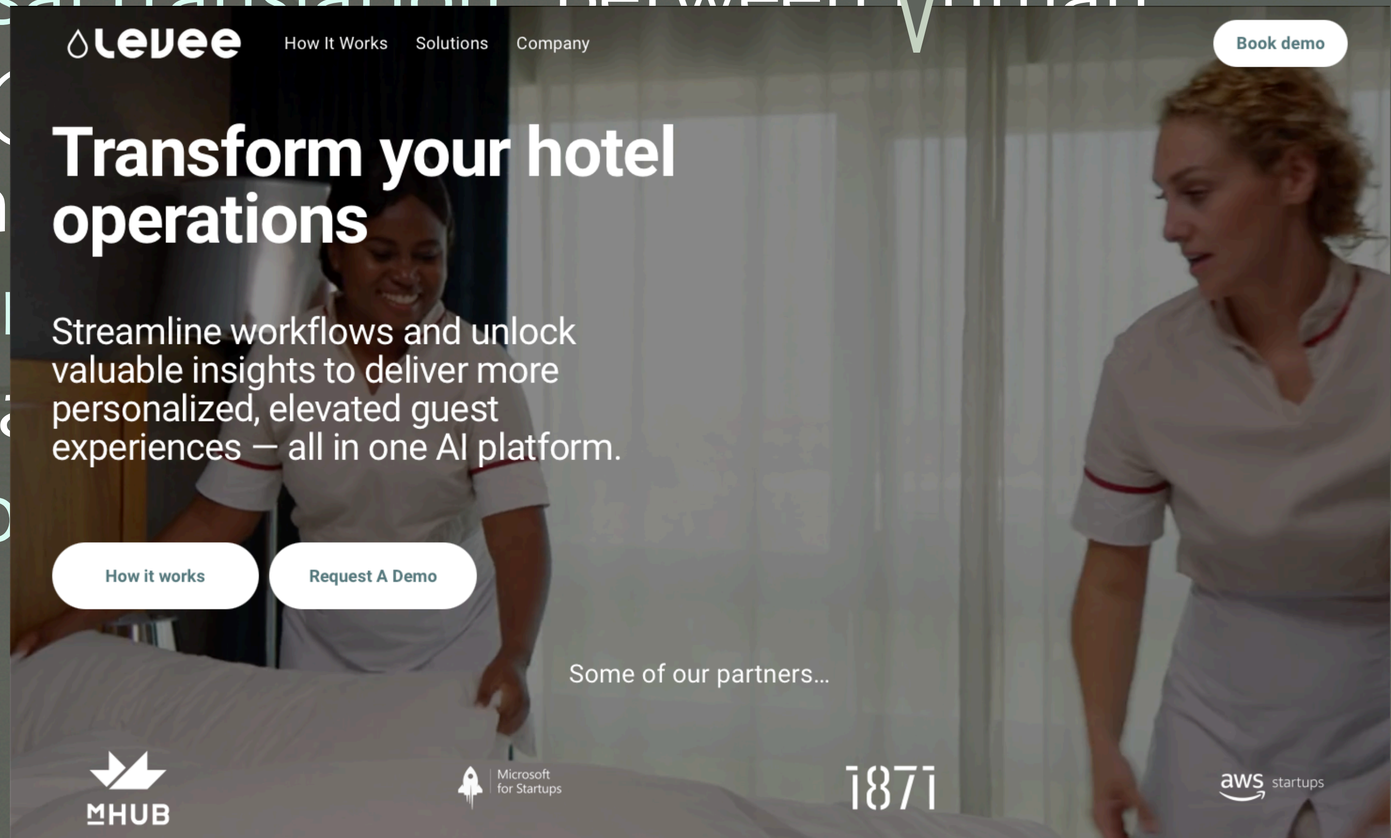
Emphasize Augmentation

- Use design patterns like RAG and Agents.
  - Combine tools with complementary strengths
  - Use models for "universal translation" between human and "other languages" (e.g., SQL).
  - Leverage relational, graph, or other data stores.
  - Use deterministic systems (templates, planning and reasoning engines) for accuracy and logic.
  - Keep humans in the loop.

Emphasize Augmentation

- Use design patterns like RAG and Agents.
  - Combine tools with complementary stren
  - Use models for "universal translation" between human and "other languages" (
  - Leverage relational, gra
  - Use deterministic syste reasoning engines) for a
  - Keep humans in the loo

An example from this year's cohort:



Regulations and Policy

Safety Concerns

- 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:
  - you acquire the information and
  - quote it with 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 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

- TOC for Gen AI inference much higher 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
    - 𝒙 Much higher costs, carbon footprint
  - Small models:
    - 𝒙 Less generalizable
    - ✓ Easy to tune to be very good for specific applications
    - ✓ Much lower costs, carbon footprint

One Solution: Smaller Models

- Mixture of Experts
  - Several smaller, cheaper models combine to match performance of one large model
- Better, easier ways to "tune" models
  - and combine them application patterns like "RAG", etc.

Few organizations need to train models from scratch. Instead, they should start with good, "open" models and tune them for their needs.

Generative AI
in Five Years?

What Problems Are Temporary?

Hardware and energy costs will plummet

- New, more efficient accelerator architectures ("Post GPU")
- New, more efficient model architectures ("Post Transformer")
- Much more efficient training, tuning, and inference

We will know what really does and doesn't work

- Probabilistic models will always hallucinate…
- … so we'll combine tools

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 lingering concerns about safety, jobs, …

A revival of writing, painting, photography, …

- We'll be sick of AI-generated content

Thank You!

- Visit thealliance.ai
- Let me know what you think!
  - dean.wampler@ibm.com
  - Mastodon and Bluesky: @deanwampler
  - Other talks: deanwampler.com/talks

Extra Slides

Notes

© 2004-2024, Dean Wampler, except where noted. The photos are based on my photographs (flickr.com/photos/deanwampler/), but all 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 the other images, I used Firefly to add elements to my image.

1. Cover and end slide images were both generated by Firefly using the following sunset image as a reference image, which taken from Clingmans Dome, Great Smoky Mountains NP: flickr.com/photos/deanwampler/51664228468/in/album-72157720120215384/
2. "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/
3. "Alignment", an Oregon coast image enhanced with Firefly: flickr.com/photos/deanwampler/4850305877/in/album-72157624506732831/
4. "Regulation and Policy", a fake city hall or parliament building where I used a night-time image of the Brussels City Hall as the reference image (not on Flickr).
5. "Total Cost of Ownership", a Chicago Park image enhanced with Firefly: flickr.com/photos/deanwampler/53419199087/in/dateposted-public/
6. "Developer and End User Education", image was generated by Firefly using the Chicago Park image as a reference image shown in the "Total Cost of Ownership" section. flickr.com/photos/deanwampler/53419199087/in/dateposted-public/

The **AI Alliance**

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

Founding Members and Collaborators*

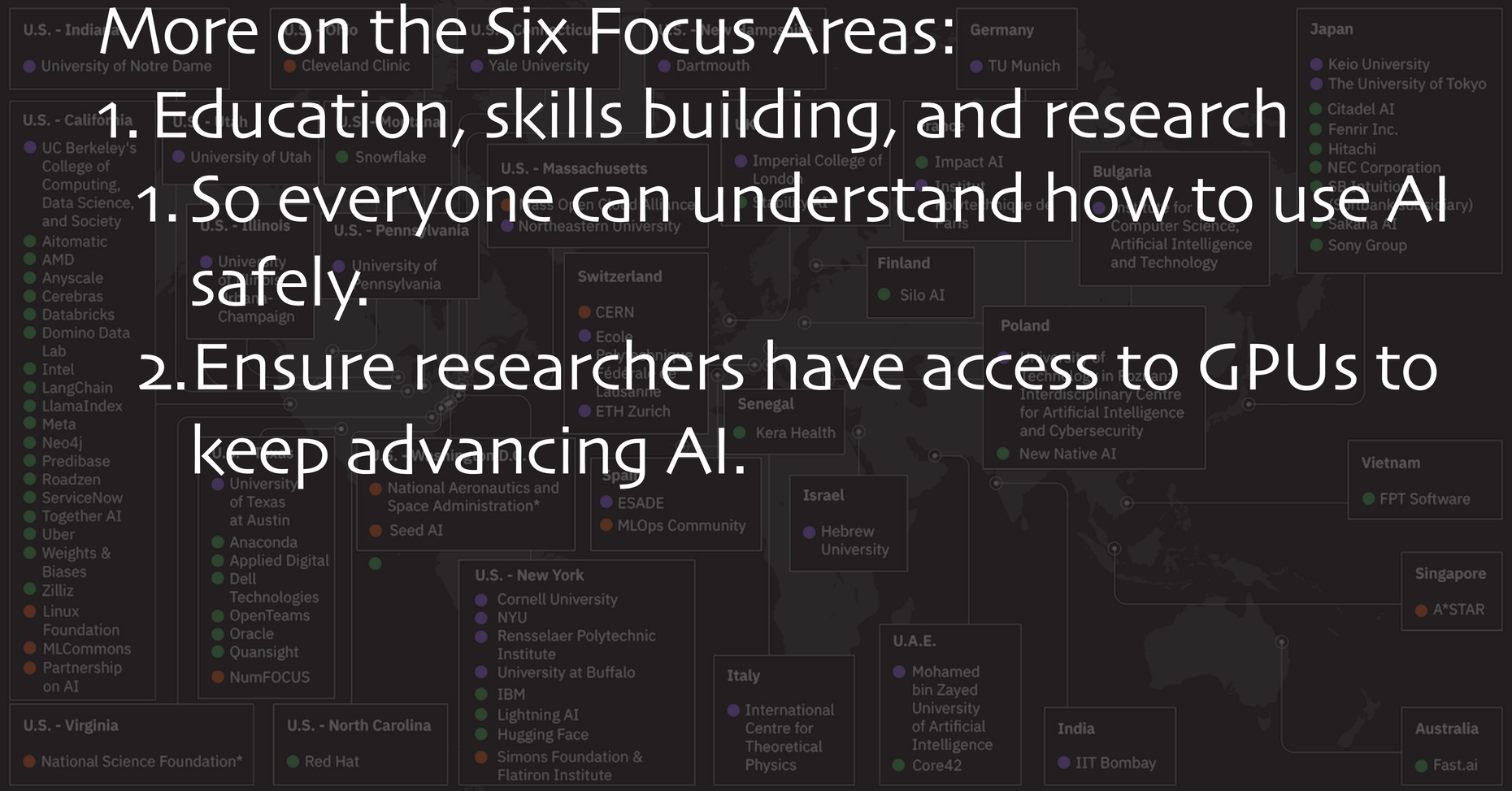
- Universities
- Startups & Enterprises
- Science Organizations & Non-profits

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

The **AI Alliance**

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



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.

The **AI Alliance**

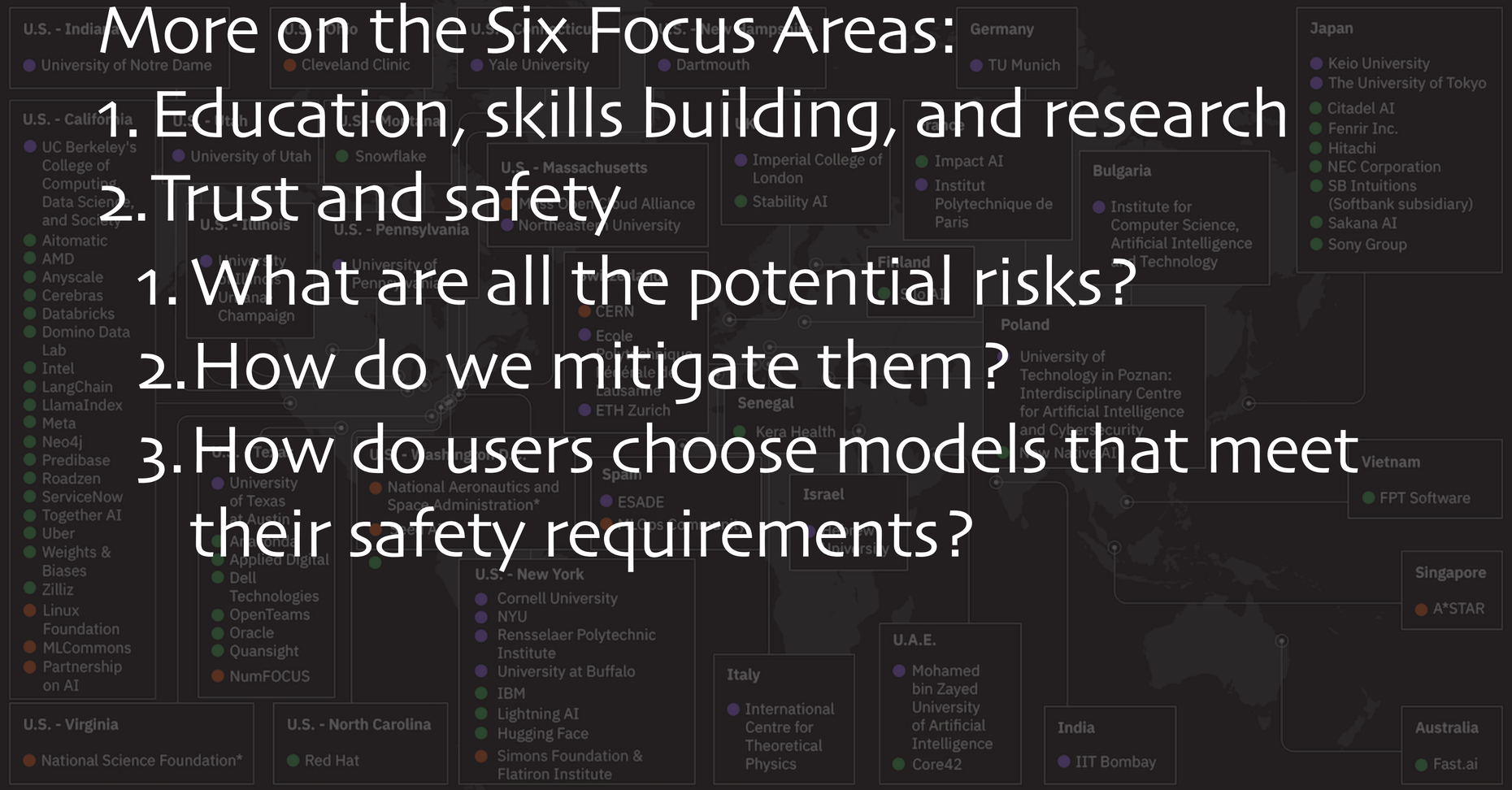
A community of technology creators, developers and adopters collaborating to advance safe, responsible AI rooted in open innovation.

Founding Members and Collaborators*

- Universities
- Startups & Enterprises
- Science Organizations & Non-profits

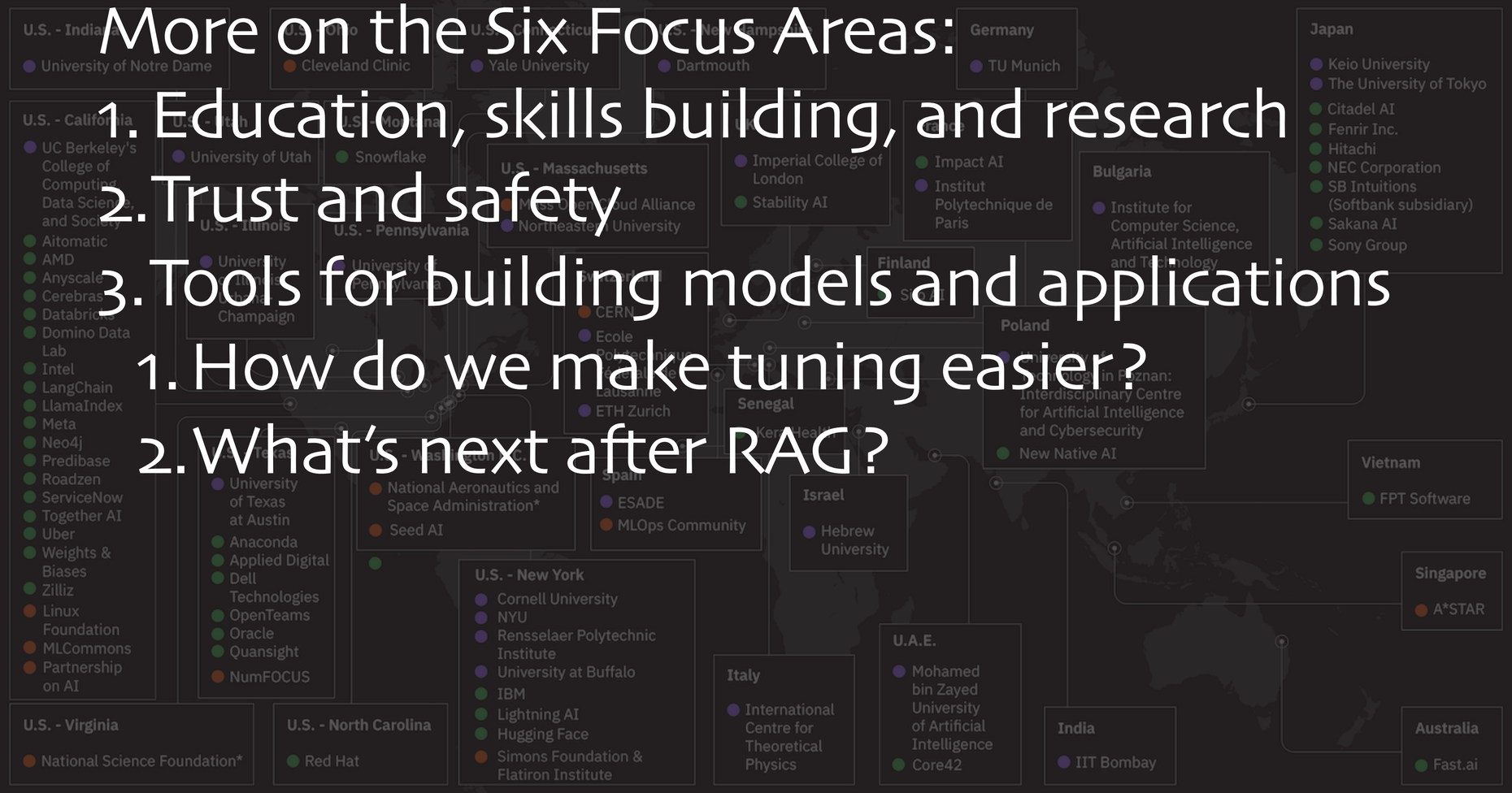
More on the Six Focus Areas:

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?



The **AI Alliance**

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



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?

The **AI Alliance**

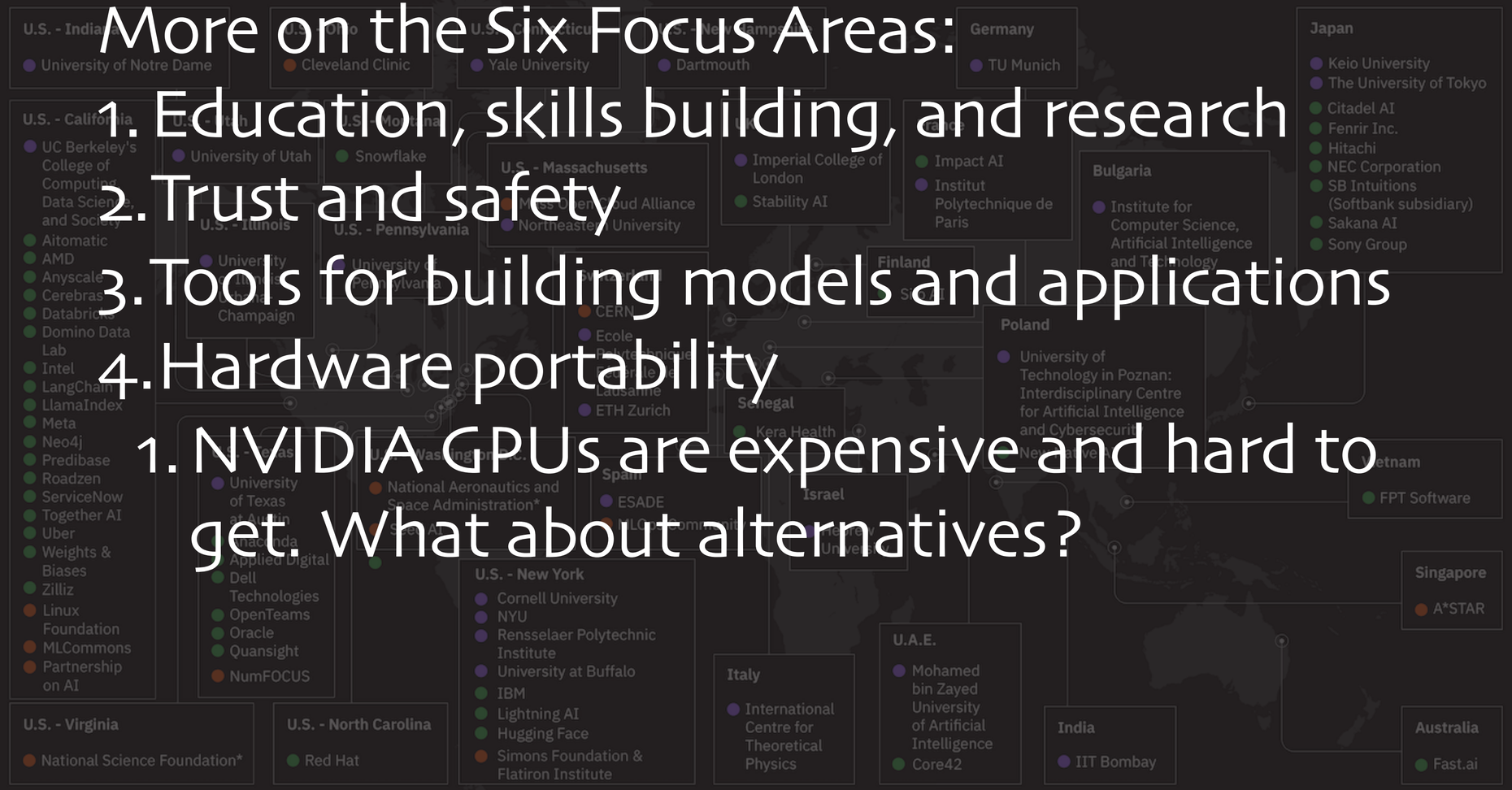
A community of technology creators, developers and adopters collaborating to advance safe, responsible AI rooted in open innovation.

Founding Members and Collaborators*

- Universities
- Startups & Enterprises
- Science Organizations & Non-profits

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. What about alternatives?



The **AI Alliance**

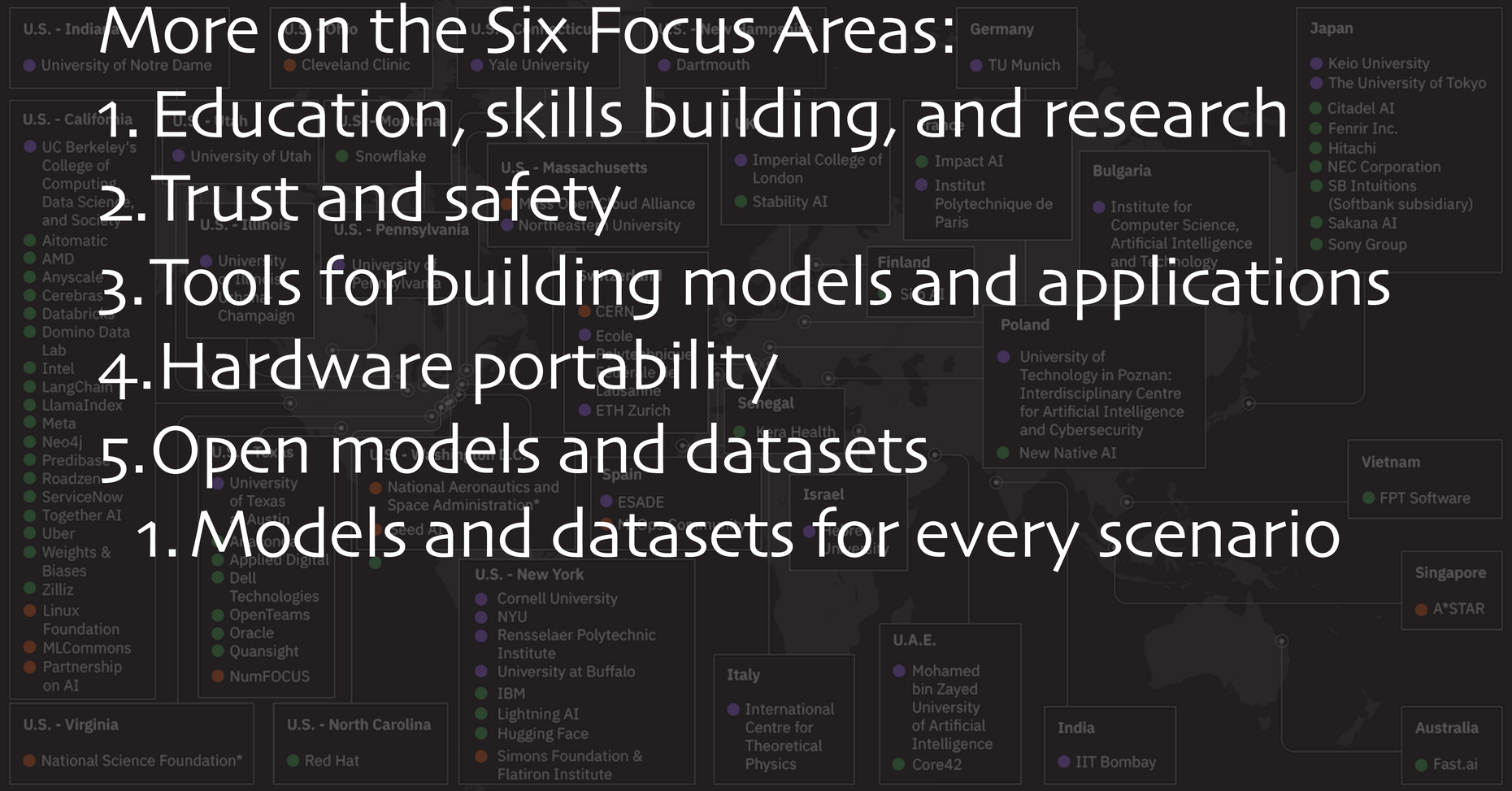
A community of technology creators, developers and adopters collaborating to advance safe, responsible AI rooted in open innovation.

Founding Members and Collaborators*

- Universities
- Startups & Enterprises
- Science Organizations & Non-profits

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

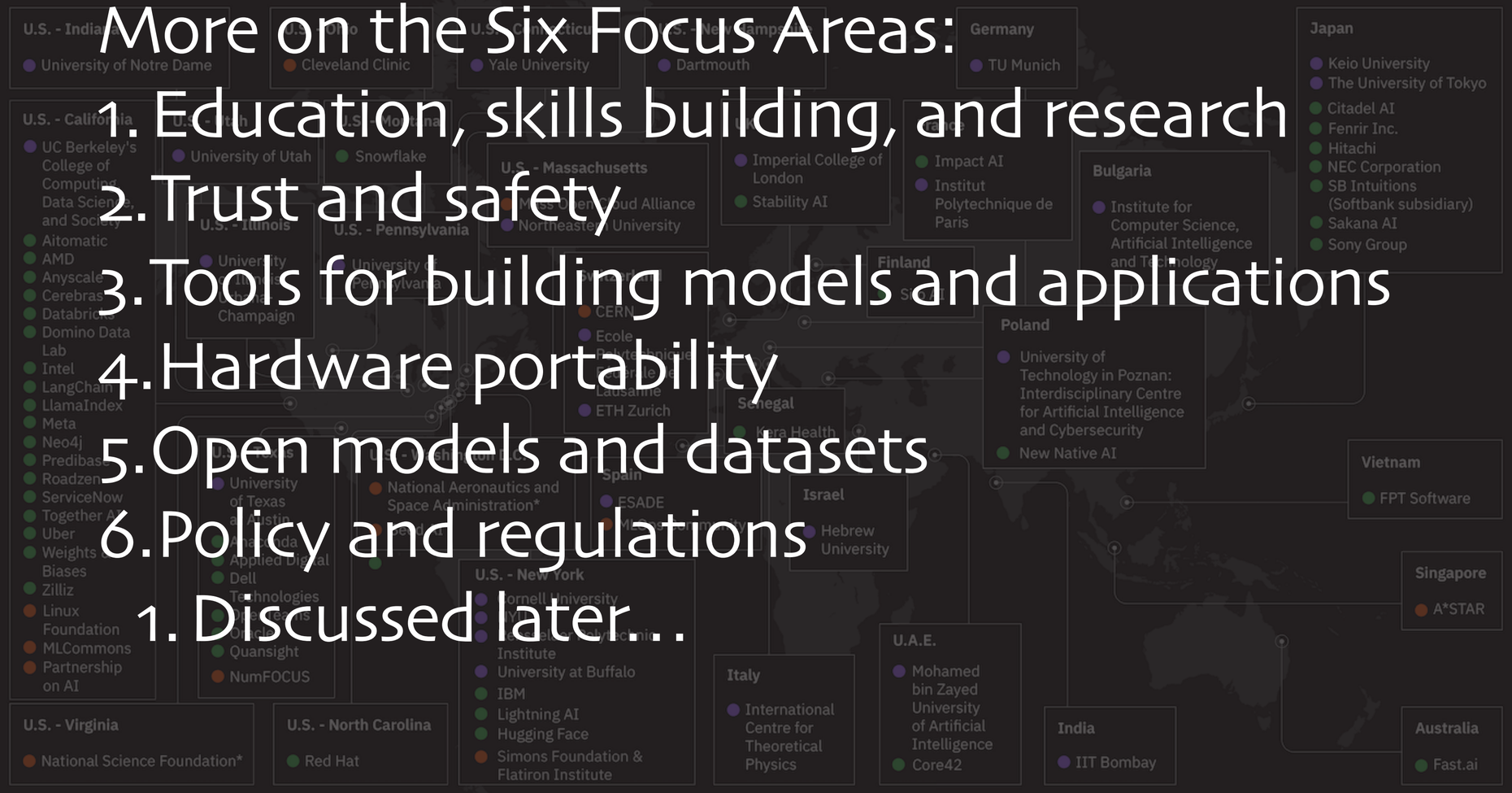


The **AI Alliance**

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

Founding Members and Collaborators*

- Universities
- Startups & Enterprises
- Science Organizations & Non-profits



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…