Understanding the AI ecosystem

Day 2. What everyone is doing

Understanding the AI ecosystem

- State of AI
- Current work in AI governance
- Main stakeholders and their POV on AI safety

State of AI

And its transformative impact

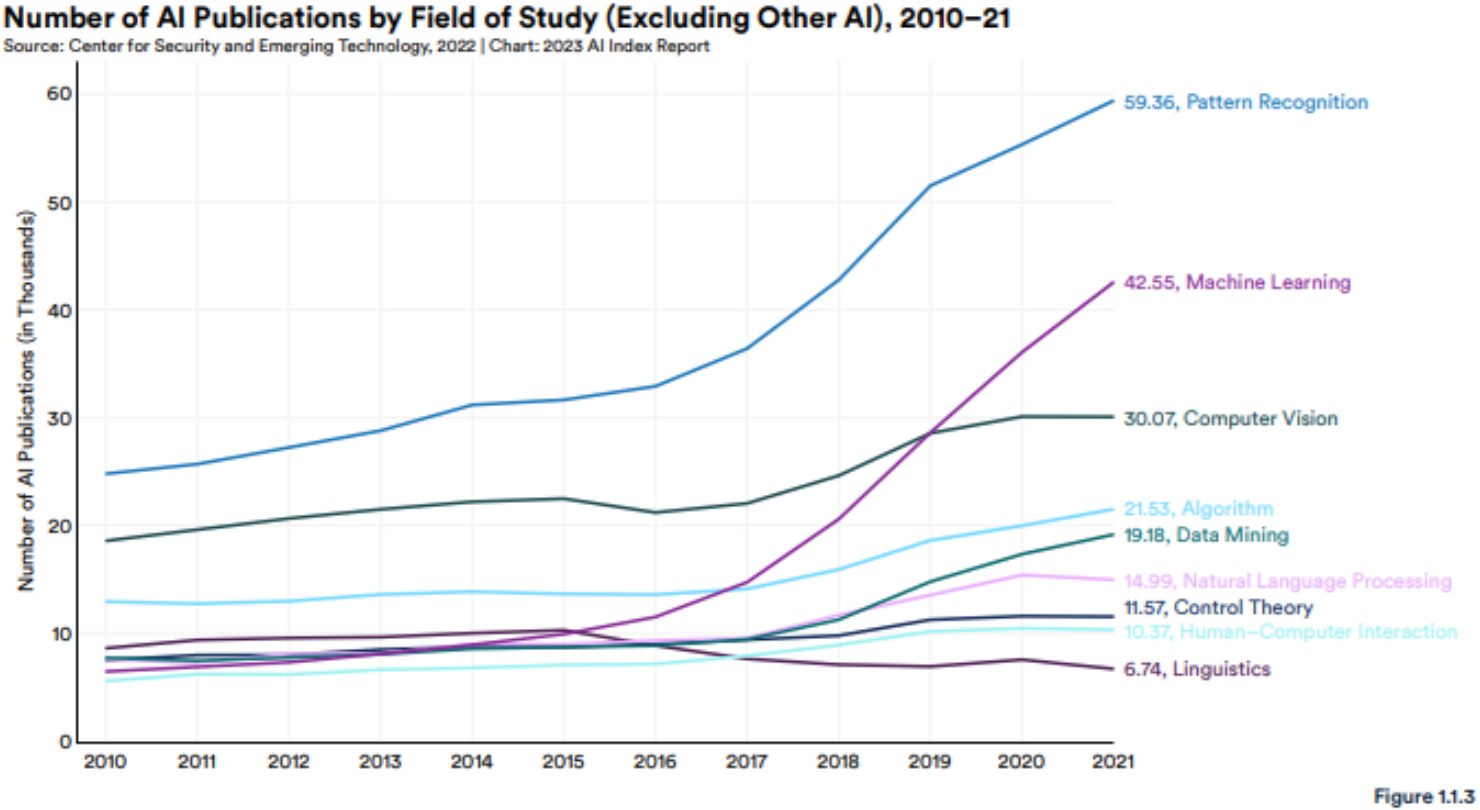
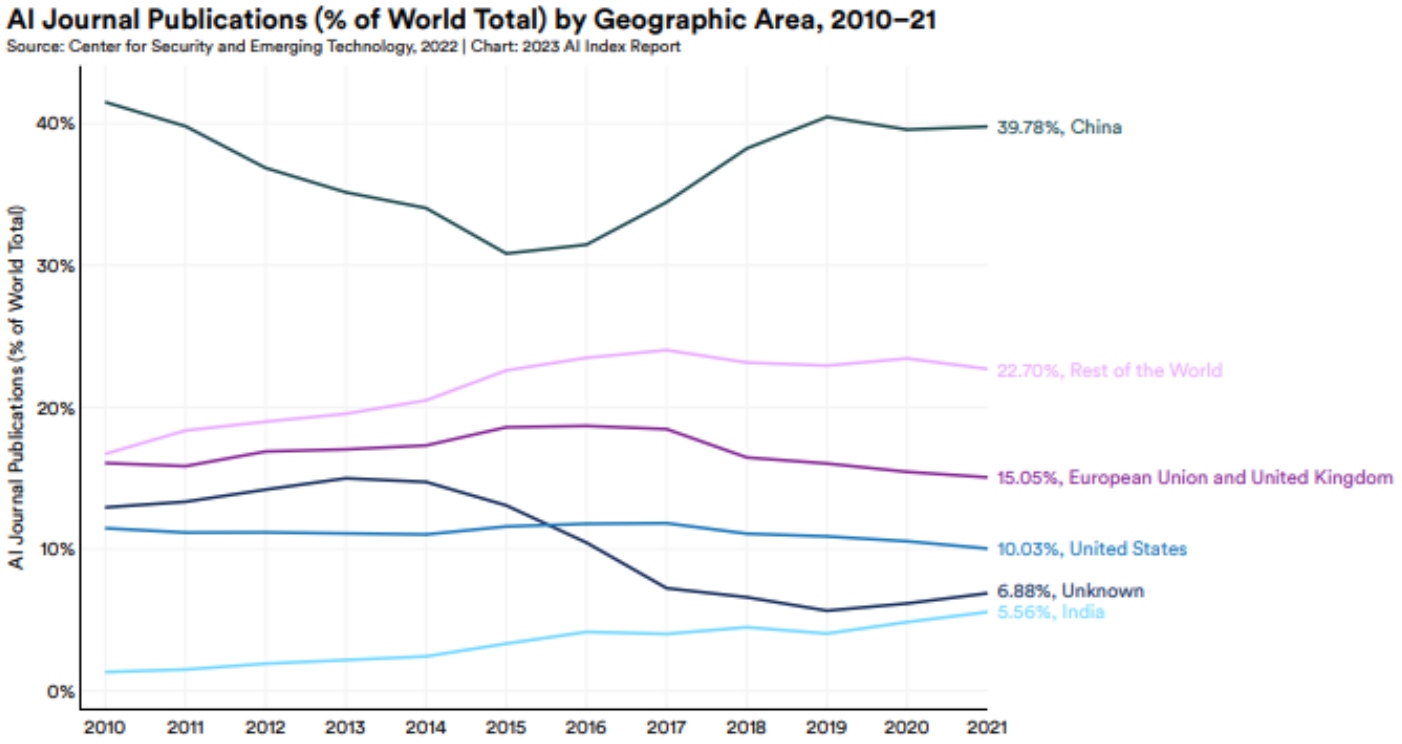
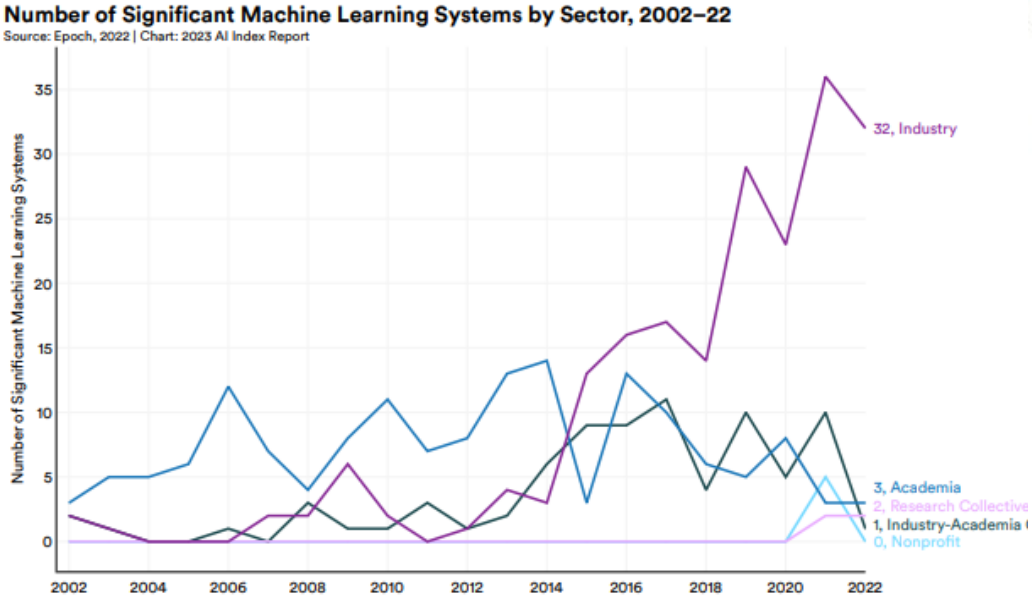
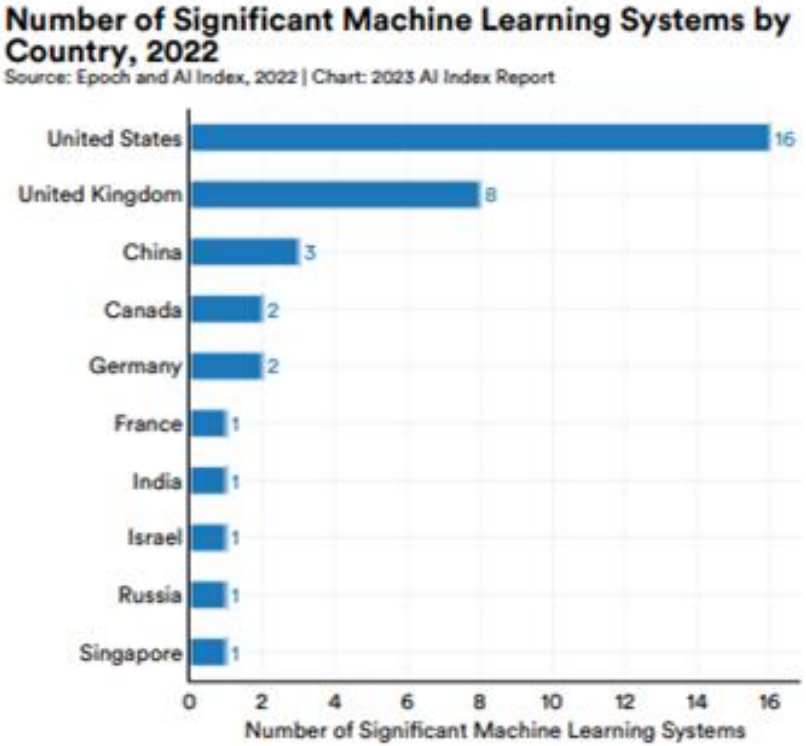

Number of AI Publications by Field of Study (Excluding Other AI), 2010–21
Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report
Number of AI Publications (in Thousands)
60
50
40
30
20
10
0
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
59.36, Pattern Recognition
42.55, Machine Learning
30.07, Computer Vision
21.53, Algorithm
19.18, Data Mining
14.99, Natural Language Processing
11.57, Control Theory
10.37, Human–Computer Interaction
6.74, Linguistics

Figure 1.1.3

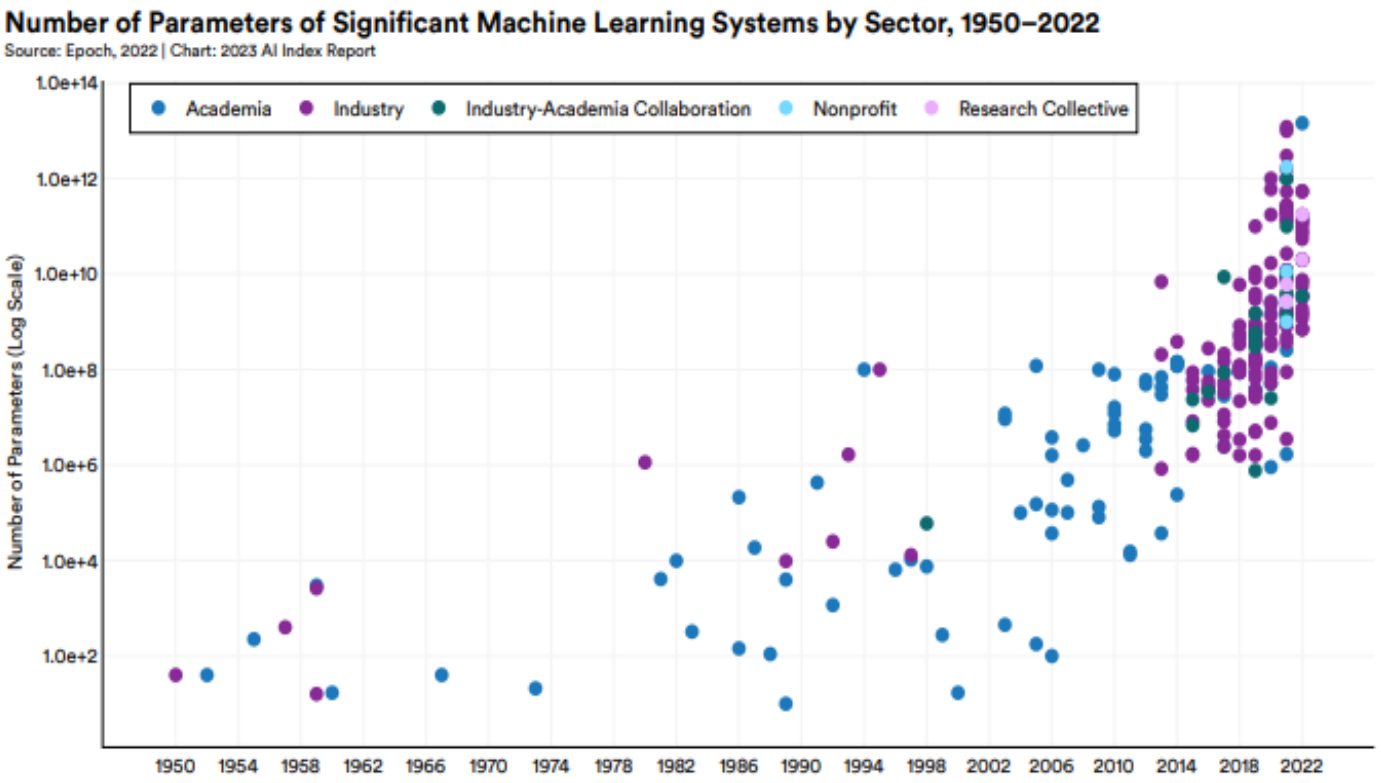


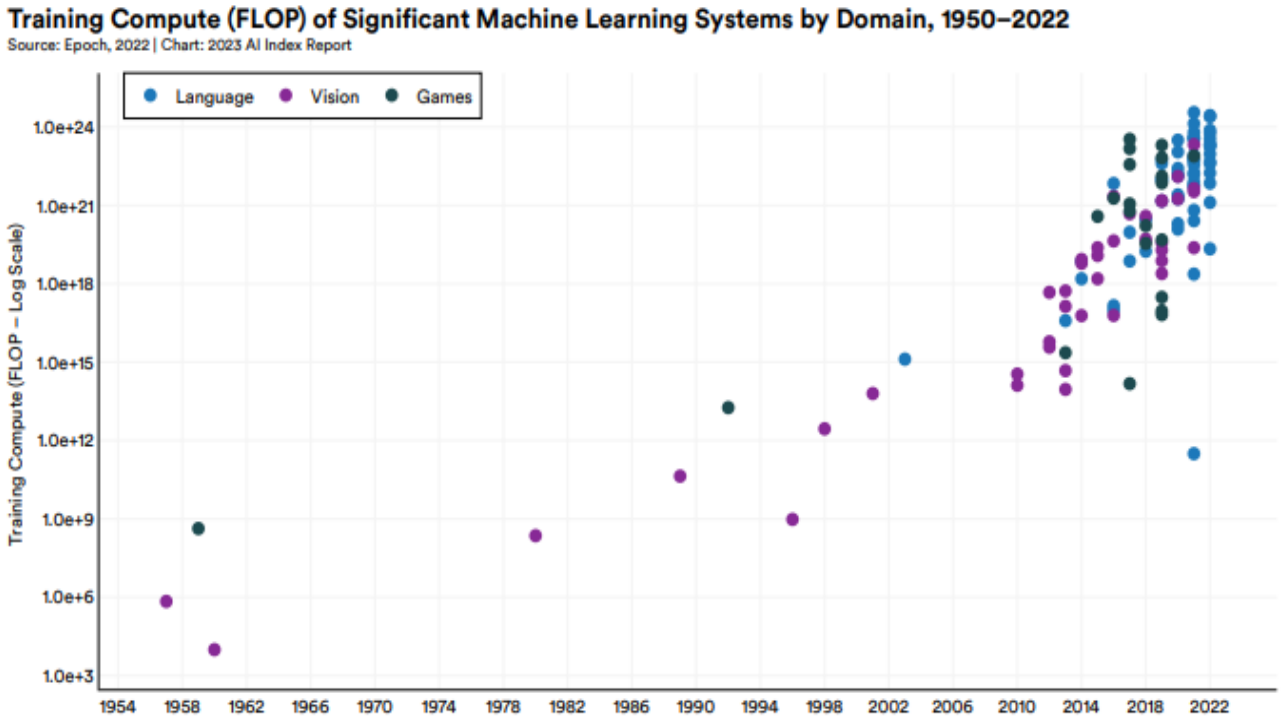
State of AI > Research

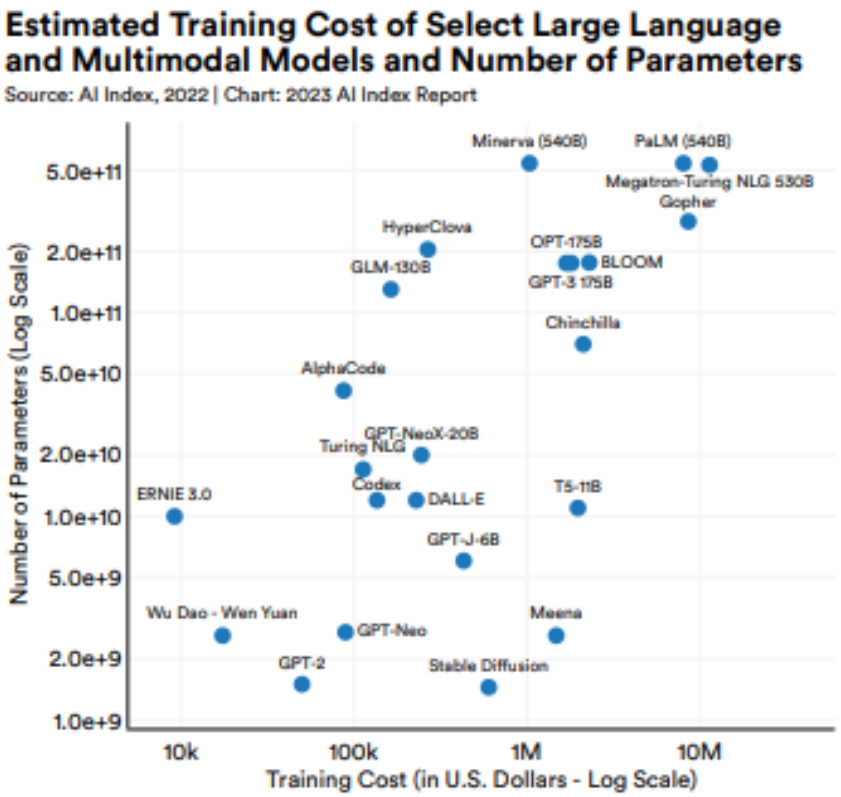


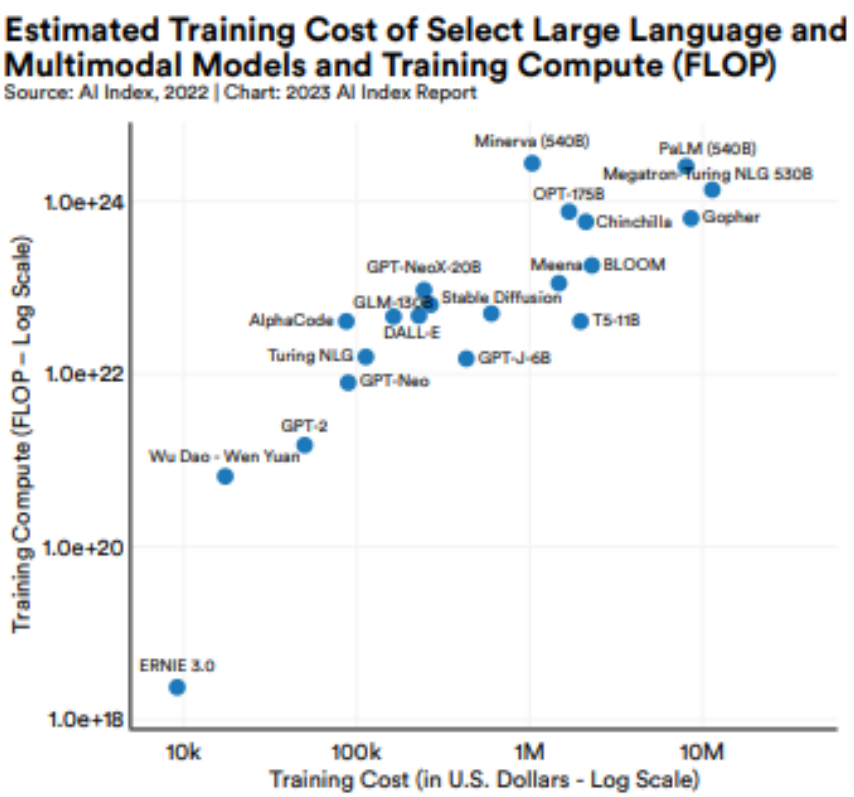


State of AI > Research


Number of Parameters of Significant Machine Learning Systems by Sector, 1950–2022
Source: Epoch, 2022 | Chart: 2023 AI Index Report
Academia
Industry
Industry-Academia Collaboration
Nonprofit
Research Collective
Number of Parameters (Log Scale)
1.0e+14
1.0e+12
1.0e+10
1.0e+8
1.0e+6
1.0e+4
1.0e+2
1950
1954
1958
1962
1966
1970
1974
1978
1982
1986
1990
1994
1998
2002
2006
2010
2014
2018
2022


Training Compute (FLOP) of Significant Machine Learning Systems by Domain, 1950–2022
Source: Epoch, 2022 | Chart: 2023 AI Index Report
Language
Vision
Games
Training Compute (FLOP – Log Scale)
1.0e+24
1.0e+21
1.0e+18
1.0e+15
1.0e+12
1.0e+9
1.0e+6
1.0e+3
1954
1958
1962
1966
1970
1974
1978
1982
1986
1990
1994
1998
2002
2006
2010
2014
2018
2022


Estimated Training Cost of Select Large Language and Multimodal Models and Number of Parameters
Source: AI Index, 2022 | Chart: 2023 AI Index Report
Number of Parameters (Log Scale)
5.0e+11
2.0e+11
1.0e+11
5.0e+10
2.0e+10
1.0e+10
5.0e+9
2.0e+9
1.0e+9
10k
100k
1M
10M
Training Cost (in U.S. Dollars - Log Scale)
Minerva (540B)
PaLM (540B)
Megatron-Turing NLG 530B
Gopher
HyperClova
OPT-175B
BLOOM
GLM-130B
GPT-3 175B
Chinchilla
AlphaCode
GPT-NeoX-20B
Turing NLG
Codex
DALL-E
T5-11B
ERNIE 3.0
GPT-J-6B
Wu Dao - Wen Yuan
GPT-Neo
Meena
GPT-2
Stable Diffusion


Estimated Training Cost of Select Large Language and Multimodal Models and Training Compute (FLOP)
Source: AI Index, 2022 | Chart: 2023 AI Index Report
Training Compute (FLOP – Log Scale)
1.0e+24
1.0e+22
1.0e+20
1.0e+18
10k
100k
1M
10M
Training Cost (in U.S. Dollars - Log Scale)
Minerva (540B)
PaLM (540B)
Megatron-Turing NLG 530B
OPT-175B
Chinchilla
Gopher
GPT-NeoX-20B
Meena
BLOOM
GLM-130B
Stable Diffusion
AlphaCode
DALL-E
T5-11B
Turing NLG
GPT-J-6B
GPT-Neo
GPT-2
Wu Dao - Wen Yuan
ERNIE 3.0

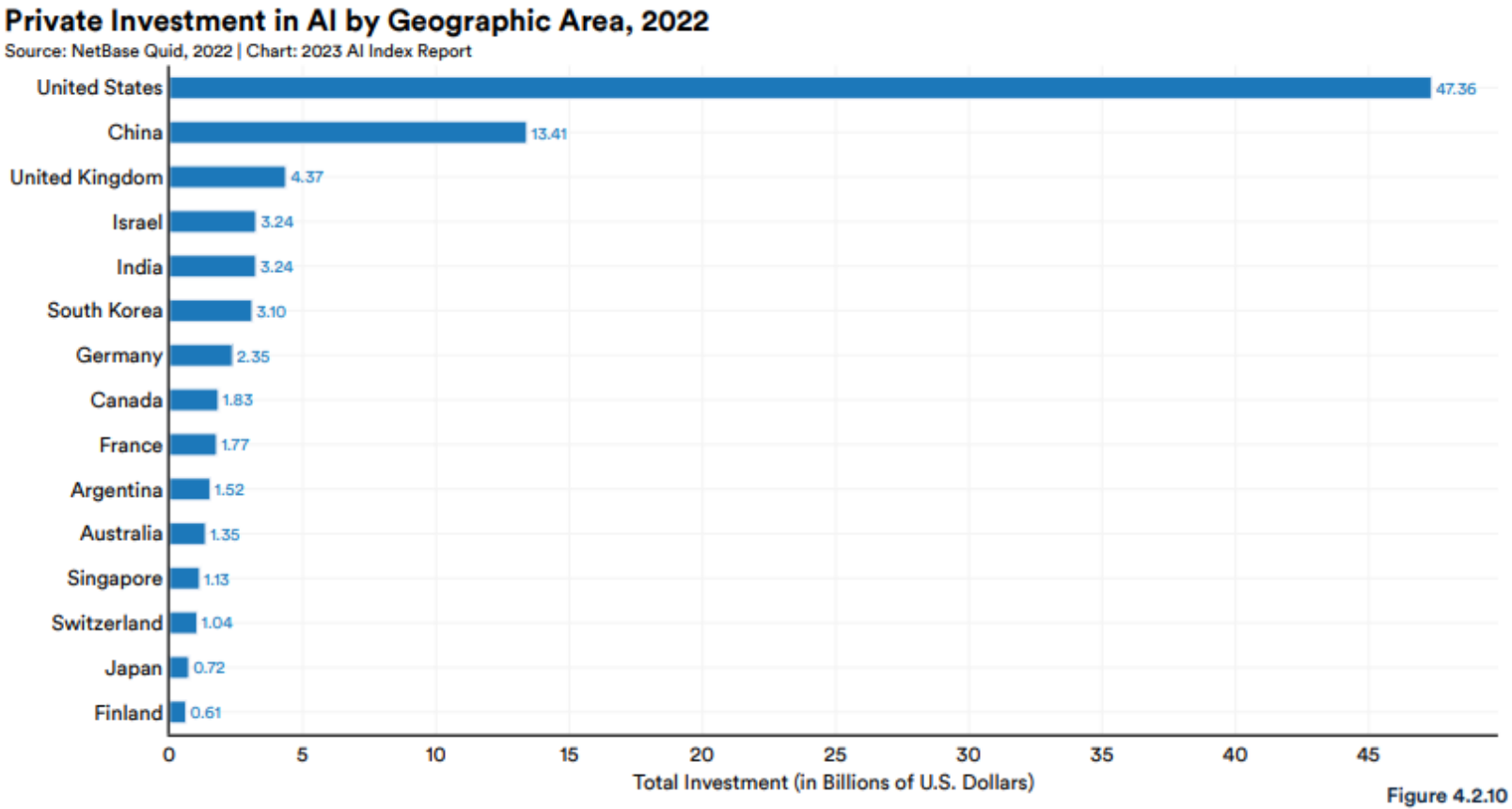
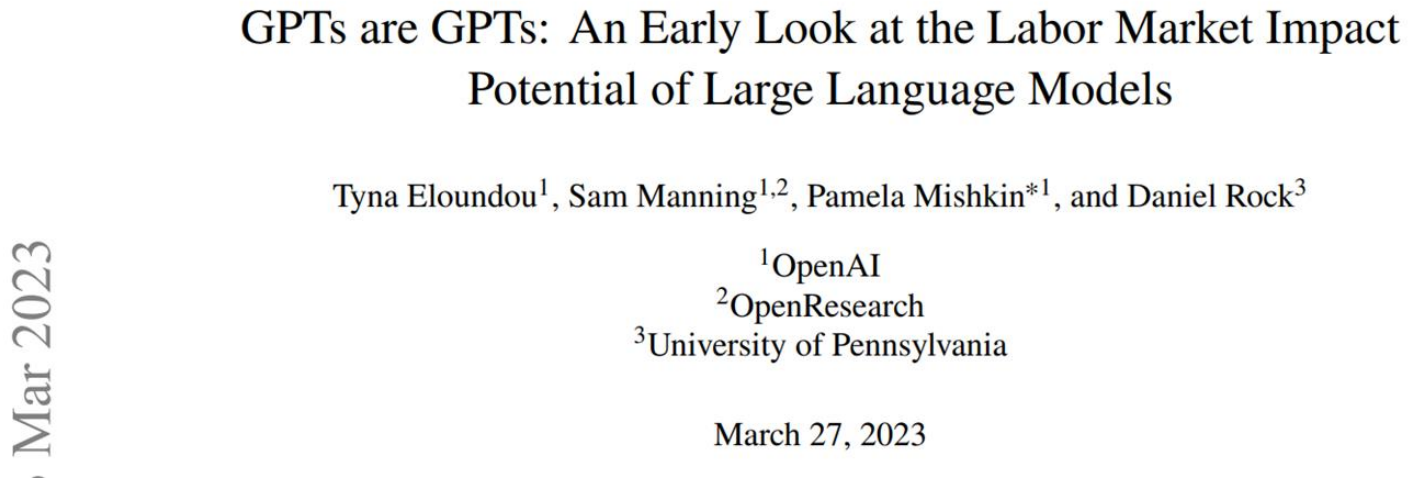

Private Investment in AI by Geographic Area, 2022
Source: NetBase Quid, 2022 | Chart: 2023 AI Index Report
United States 47.36
China 13.41
United Kingdom 4.37
Israel 3.24
India 3.24
South Korea 3.10
Germany 2.35
Canada 1.83
France 1.77
Argentina 1.52
Australia 1.35
Singapore 1.13
Switzerland 1.04
Japan 0.72
Finland 0.61
0 5 10 15 20 25 30 35 40 45
Total Investment (in Billions of U.S. Dollars)

Figure 4.2.10

Generative AI could raise global GDP by 7%

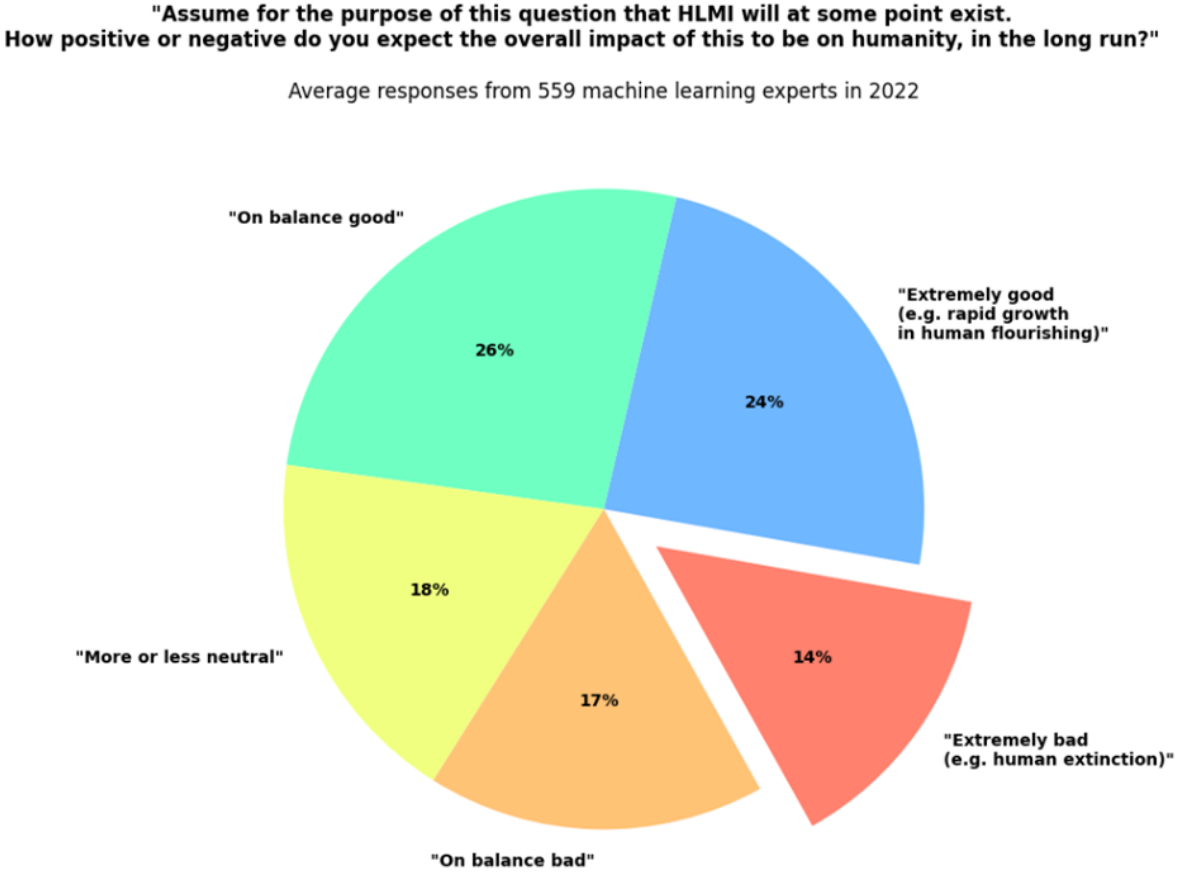


GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models

Tyna Eloundou[1], Sam Manning[1,2], Pamela Mishkin*[1], and Daniel Rock[3]

[1]OpenAI
[2]OpenResearch
[3]University of Pennsylvania

March 27, 2023


"Assume for the purpose of this question that HLMI will at some point exist.
How positive or negative do you expect the overall impact of this to be on humanity, in the long run?"
Average responses from 559 machine learning experts in 2022
"On balance good"
26%
"Extremely good
(e.g. rapid growth
in human flourishing)"
24%
18%
"More or less neutral"
14%
17%
"Extremely bad
(e.g. human extinction)"
"On balance bad"

Extinction from AI

Participants were asked:

What probability do you put on future AI advances causing human extinction or similarly permanent and severe disempowerment of the human species?

Answers

Median 5%.

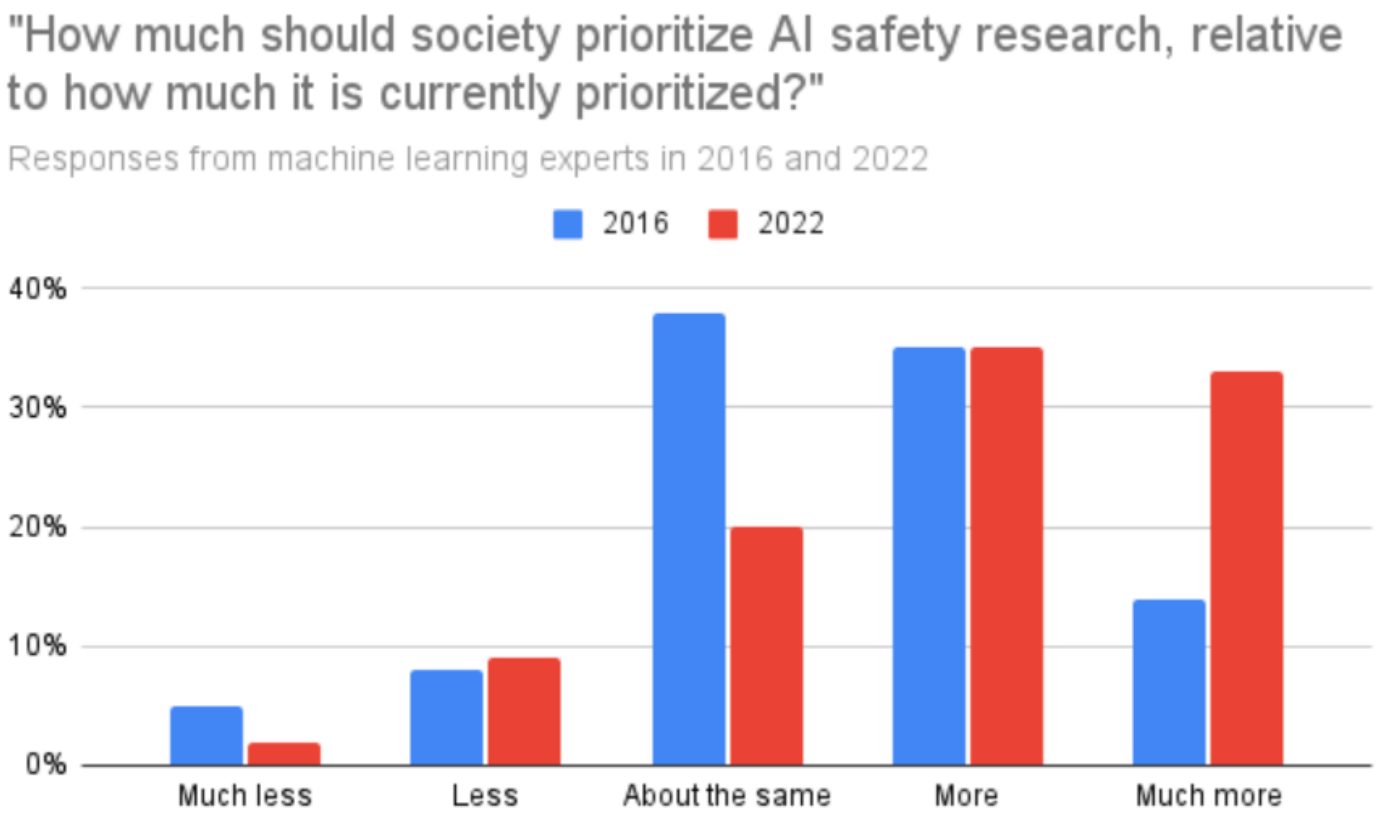
Extinction from human failure to control AI

Participants were asked:

What probability do you put on human inability to control future advanced AI systems causing human extinction or similarly permanent and severe disempowerment of the human species?

Answers

Median 10%.

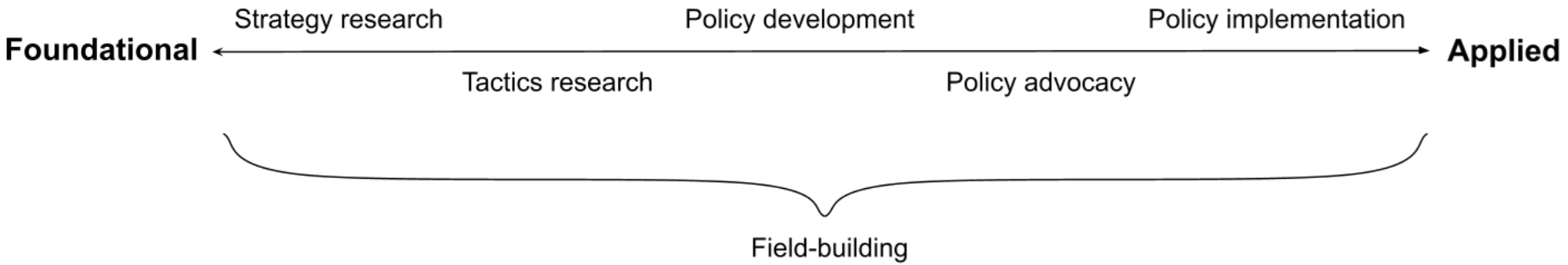


State of AI > current opinions of AI experts

Current work in AI governance

A summary

Governance work spectrum



Strategy > Strategy research

Some significant questions :

- What are the primary sources of existential risk?
- What are the AI capabilities of China? How likely is China to become an AI superpower?
- Will there be significant military interest in AI technologies? Will this lead to military AI megaprojects?

Who is working on this :

Open Philanthropy

RETHINK PRIORITIES

Legal Priorities Project

Centre for the Governance of AI

GLOBAL PRIORITIES INSTITUTE

Strategy > Surveys

Foundation for strategic decisions :

Expert opinions
- ➔ AI timelines
- ➔ Risk assessment

Public opinions
- ➔ Overton window

Who is working on this :

AI IMPACTS

RETHINK PRIORITIES

Centre for the Governance of AI

Strategy > Forecasting

Some significant questions :

- When will AGI be developed?
- Will AI takeoff be fast or slow?
- What impacts of AI should we expect on democracy or international stability in the coming years?
- Will data be a serious bottleneck for increasing the size of future AI models?

Who is working on this :

AI IMPACTS

Open Philanthropy

EPOCH

Metaculus

Group discussion !

Will data be a serious bottleneck for increasing the size of future AI models?

Industry focused governance

- ➔ Very little existing government regulation of AI currently exists
  Most important decisions about training and deployment are almost entirely made within the industry.
- ➔ Incredibly concentrated AI industry
- ➔ Possible to influence key decisions by working with a small number of actors.

Industry focused governance

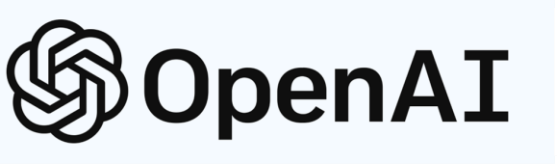
> Improving corporate decisions

Decisions on development, deployment and investment :

- DeepMind’s decision to stop releasing their models
- Decisions by OpenAI, such as not open-sourcing GPT-3, waiting six months before releasing GPT-4, giving Microsoft early-access to GPT-4
- Meta AI giving academics access to Llama

Who is working on this :

ANTHROP\C



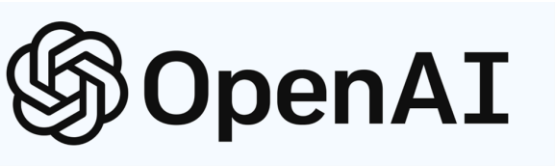
Google DeepMind

Industry focused governance

> Improving corporate structures

- Implementing good risk management procedures at AI labs
- Legal status of AI labs, e.g. capped profit (OpenAI)
- Red-teaming, evaluations, oversight boards

Who is working on this :



Google DeepMind

Industry focused governance

> Model evaluations

- Make capabilities and alignment of models measurable
- Predict when models will get dangerous

➔ Credible commitments, stronger arguments for dangers

Who is working on this :

OpenAI

ARC Evals

Industry focused governance

> Standards setting

- Dominant approach for safety in many industries
- For AI it could be :
  - Robustness & trustworthiness benchmarks
  - Bias
  - Development standards
  - Certifications and auditing standards
  - Governance standards

Who is working on this :

NIST National Institute of Standards and Technology

ISO International Organization for Standardization

CENELEC EUROPEAN COMMITTEE FOR ELECTROTECHNICAL STANDARDIZATION

future of life INSTITUTE

Industry focused governance

> Incentivizing responsible publication norms

Reduce the number of actors with access to cutting-edge AI models.

- OpenAI did not release many technical details of GPT-4
- Major releases from DeepMind has decreased in the past months.

Democratising AI: Multiple Meanings, Goals, and Methods

Elizabeth Seger, Aviv Ovadya, Ben Garfinkel, Divya Siddarth, Allan Dafoe

The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse?

Toby Shevlane, Allan Dafoe

Who is working on this :

OpenAI

Centre for the Governance of AI

Google DeepMind

Group discussion !

How careful should we be when publishing AI research ?

Government focused governance

> Legislative action

- EU AI act
- UK : first major global summit on AI safety
- US : hearing on AI in Senate
- China : national plans for AI regulation

Who is working on this :

Centre for the Governance of AI

THE FUTURE SOCIETY

THE CENTRE FOR LONG-TERM RESILIENCE

future of life INSTITUTE

Government focused governance

> Compute governance

- Current AI systems require expensive and sophisticated hardware
- This hardware is produced in a complex and global supply chain
- Regulating the supply chain can enforce standards and exclude bad actors

Some ideas :

- Monitoring capabilities
- Restricting access to compute
- Hardware based mechanisms

Who is working on this :

RETHINK PRIORITIES

Centre for the Governance of AI

CSET
CENTER for SECURITY and EMERGING TECHNOLOGY

Government focused governance

> International governance

- Ban military use of AI (FLI)
- Avoid conflict
- Knowledge sharing to prevent dangerous technology races : IAEA for AI ?
- Politically neutral hub for AI research : CERN for AI ?

Who is working on this :

RETHINK PRIORITIES

Simon Institute for Longterm Governance

Centre for the Governance of AI

future of life INSTITUTE

Field building > Grantmaking

- Prioritize which work gets funded
  > shapes the field and the strategies
- Majority of work in AIS funded by private philanthropy

Who is working on this :

EA Funds

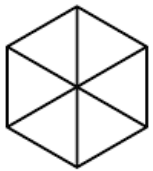
Open Philanthropy

Survival and Flourishing .Fund (SFF)

Field building > Media campaigns

- FLI's open letter
- FLI's campaign against autonomous weapons
- Coverage of AI risk in the Economist, Time, NYT
- Statement on AI Risk, Center for AI Safety
- Theory of change of the existential risk observatory

Who is working on this :



Existential Risk Observatory

Center for AI Safety

future of life INSTITUTE

Field building > Outreach

Broadening the field :

- Conferences
- Networking between industry and policy
- Events to connect different fields

> Allan Dafoe : *"I think we want to build a Metropolis---a hub with dense connections to the broader communities of computer science, social science, and policymaking---rather than an isolated Island."*

Who is working on this :

Centre for the Governance of AI

CENTRE FOR THE STUDY OF EXISTENTIAL RISK

Legal Priorities Project

future of life INSTITUTE

Field building > Scouting and training talent

Main talent pipeline :

- Interest in AI risk
- Join reading group or short bootcamp
- Test fit in one of the (fairly competitive) summer opportunities
- Join a longer fellowship
- Begin working in academia, in industry, for a think tank, or for government

Talent pipeline

- Interest in AI risk
- **Reading groups / short bootcamps**
- Test fit in Summer opportunities
- Longer fellowships
- Begin working in academia, in industry, for a think tank, or for government

What exists currently :

AI Safety Fundamentals

A project by

BlueDot Impact

ARENA

Talent pipeline

- Interest in AI risk
- Reading groups / short bootcamps
- **Test fit in Summer opportunities**
- Longer fellowships
- Begin working in academia, in industry, for a think tank, or for government

What exists currently :

ERA

Stanford | Existential Risks Initiative

Swiss Existential Risk Initiative

Talent pipeline

- Interest in AI risk
- Reading groups / short bootcamps
- Test fit in Summer opportunities
- **Longer fellowships**
- Begin working in academia, in industry, for a think tank, or for government

What exists currently :

EU Tech Policy Fellowship 2023

Open Philanthropy Technology Policy Fellowship

Centre for the Governance of AI

Winter Fellowship 2024

SERI ML Alignment Theory Scholars Program

Potential work in AI governance

Some areas that could be explored

Data governance

- Training data is usually scraped from internet
- Current legal situation for what data may and may not be used is unclear

  - DeepMind is being sued for mishandling protected medical data
  - Stable Diffusion and Midjourney are being sued by artists for producing derivative works
  - Copilot is being sued for reciting code without stating the authors

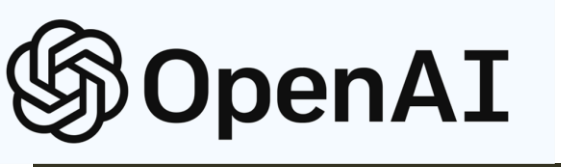
Group discussion !

- Does AI governance go against innovation ?
- How can governments and regulatory bodies keep pace with the rapid advancements in AI technology?
- What lessons can we learn from past technologies in shaping AI governance for the future?
- Think about other AI governance approaches that could be implemented !

Main stakeholders

And their POV on AI safety

Main labs > OpenAI



Our approach to alignment research

1. Training AI systems using human feedback
2. Training AI systems to assist human evaluation
3. Training AI systems to do alignment research

375 people (most in capabilities)

- Superalignment team
- Policy research team
- RL, Human Data, Security, Applied Safety, and Trust and Safety teams

“Over the next four years the division will have access to 20% of the company’s processing capacity to build a “human-level automated alignment researcher” that can be scaled up to supervise superintelligence.”

Main labs > Google Deepmind

1,567 people (most in capabilities)

- Technical safety team
- Ethics and public engagement teams

- Signed pledge against use of their technologies for lethal autonomous weapons
- Specification, robustness, and assurance

Main labs > Meta AI

Meta AI

10000+ employees

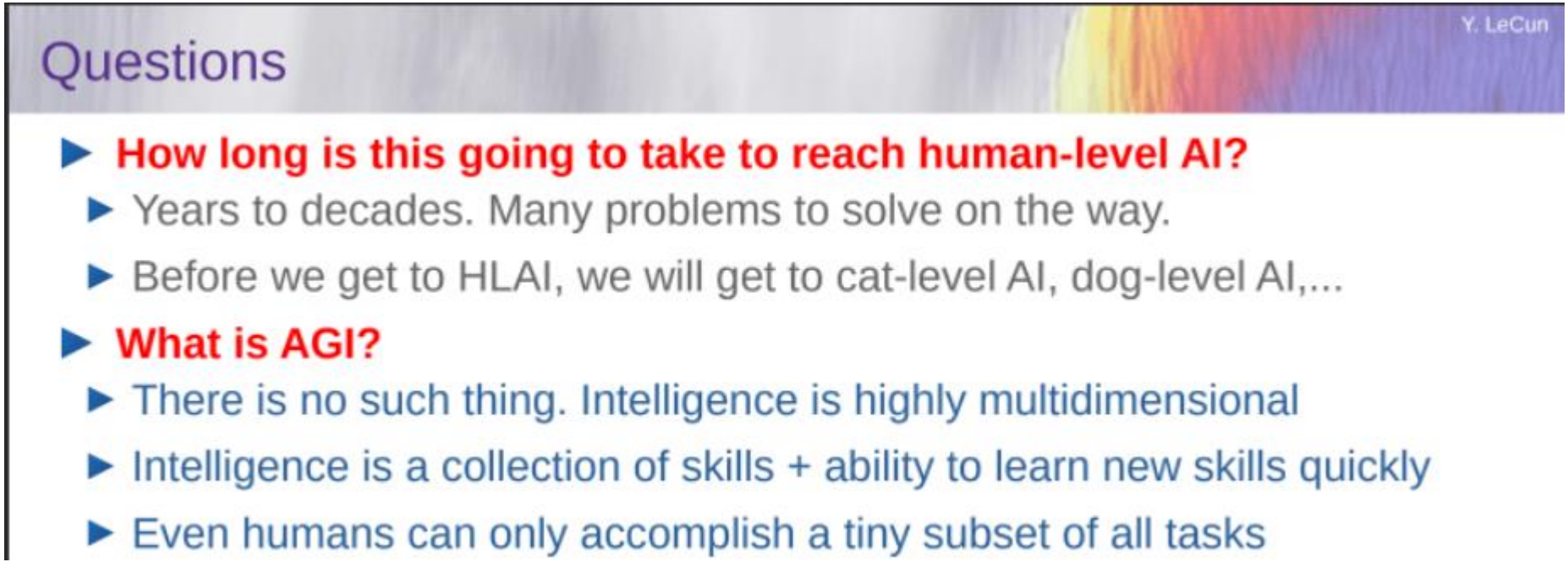
Don’t Fear the Terminator

Artificial intelligence never needed to evolve, so it didn’t develop the survival instinct that leads to the impulse to dominate others

By Anthony Zador, Yann LeCun on September 26, 2019

One of the 'godfathers' of AI says concerns the technology could pose a threat to humanity are 'preposterously ridiculous'

Main labs > Meta AI



Main labs > Meta AI

- **Are there risks associated with human-level AI?**
  - Yes, as with every technology
  - But all those risks can be mitigated
  - Disinformation, propaganda, hate, spam,...: **AI is the solution!**
- **Should AI research be open source or heavily regulated?**
  - In a future where everyone interacts with AI assistants for everything in their daily lives, the base models **must** be open.
  - Having them controlled by a small number of company is too dangerous
- **Will robots take over the world?**
  - No! this is a projection of human nature on machines
  - Intelligence is not correlated with a desire to dominate, even in humans
  - Objective-Driven AI systems will be made subservient to humans

Main labs > Meta AI

- **How to solve the alignment problem?**
  - Through trial and error and testing in sand-boxed systems
  - We are very familiar with designing objectives for human and superhuman entities. It's called law making.
  - What if bad people get their hand on on powerful AI? Their evil AI will be inferior to the Good Guys' AI police.

- **What are the benefits of human-level AI?**
  - AI will amplify human intelligence
  - Everyone will have a staff of intelligent agents working for them

- **AI will bring a new era of enlightenment, a renaissance to humanity**

Group discussion !

- Do you agree with Yann Lecun’s POV ?

“Artificial intelligence never needed to evolve, so it didn’t develop the survival instinct that leads to the impulse to dominate others”

Main actors > EU

THE AI ACT

Unacceptable Risk

Highest level of risk prohibited in the EU. Includes AI systems using e.g. subliminal manipulation or general social scoring.

High Risk

Most regulated AI systems, as these have the potential to cause significant harm if they fail or are misused, e.g. if used in law enforcement or recruiting.

Limited Risk

Includes AI systems with a risk of manipulation or deceit, e.g. chatbots or emotion recognition systems. Humans must be informed about their interaction with the AI.

Minimal Risk

All other AI systems, e.g. a spam filter, which can be deployed without additional restrictions.

EU AI Act risk framework, 2022

The Brussels Effect and Artificial Intelligence:

How EU regulation will impact the global AI market

Charlotte Siegmann* and Markus Anderljung* | August 2022

Main actors > China

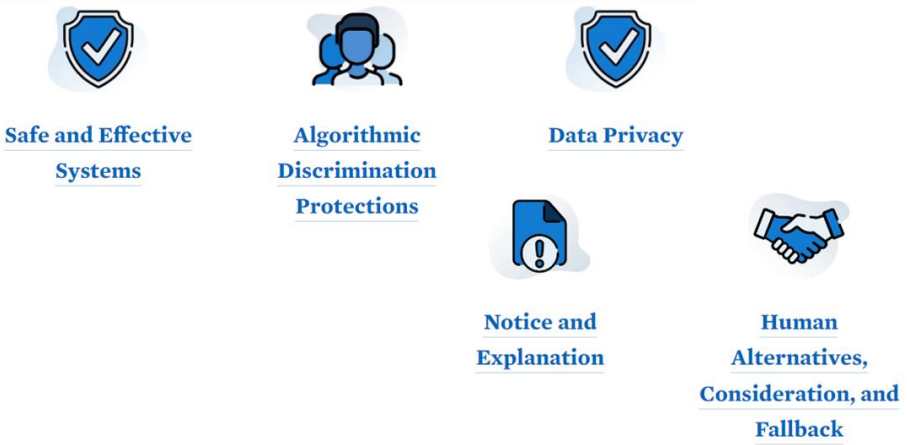
Regulations target recommendation algorithms for disseminating content, synthetically generated images and video, and generative AI systems

➔ China is the largest producer of AI research in the world, and its regulations will drive new research as companies seek out techniques to meet regulatory demands.

International discourse on Chinese AI governance often biased

- Dismissed as irrelevant
- Used as political prop to advance other agendas (US)

Main actors > US



"Systems should undergo **pre-deployment testing, risk identification and mitigation**, and **ongoing monitoring** that demonstrate they are safe and effective based on their intended use, **mitigation of unsafe outcomes** including those beyond the intended use, and adherence to domain-specific standards.
Outcomes of these protective measures should include the **possibility of not deploying the system or removing a system from use**."

Questions !