Introduction to AI Safety, Ethics, and Society

Governance

Roadmap

1. Landscape: Which actors and tools can govern AI?
2. Growth: How will AI impact economic growth?
3. Distribution: How will costs, access to and power of AI be distributed?
4. Corporate governance: How can corporations be effectively governed?
5. National governance: How can national governments promote safety?
6. International governance: How can we govern AIs internationally?
7. Compute governance: Does governing compute help govern AIs?

The Landscape: Actors (1/2)

Companies develop and deploy AIs, typically seeking profits.

- OpenAI, xAI, Anthropic, Meta, Google DeepMind
- Future companies could emerge like AI insurance companies

Nonprofits/civil society serve a variety of purposes, from developing safe AIs, conducting technical research, facilitating collaborations, and more.

- Center for AI Safety, Future of Life Institute, MIRI, WEF, RAND
- Future nonprofits could provide IDs to AI agents

National governments can influence the development and deployment of AIs

The Landscape: Actors (2/2)

International alliances help coordinate efforts between countries.

- NATO, G7, G20, Five Eyes (AUS/CAN/NZ/UK/US), EU, UN

Individuals influencing AI

- Musk, Altman, Hinton/Bengio/LeCun, Yudkowsky

The Landscape: Tools

Information affects how people think and make decisions.

- maintain AI chip registry, query companies with Defense Production Act

Financial (dis)incentives motivate corporate and individual behavior.

- taxes, liability, export controls/advanced market commitments, government procurement contracts

Standards set expectations for behavior, while **regulations** and **laws** directly enforce behavior.

- NIST, ISO, and others develop standards; state legislatures and US congress set laws

Growth

How will AI impact economic growth?

Explosive economic growth

Some researchers argue that AI-driven automation could enable unprecedented economic growth, up to ten times the current growth rate, doubling global output every few years.

AI could lead to an explosion in the effective economic population:

- Positive feedback loop between population and productivity.
- AIs substituting for human labor would lift a major bottleneck on productivity.
- Growth of AIs, both count and capabilities, can be faster than humans.

Potential bottlenecks on AI automation

Physical constraints and non -accumulable factors e.g. land, energy

Regulations and human preferences for social interaction

Challenges in accelerating R&D if "low-hanging fruit" is already taken

Bottlenecks from non -automated tasks e.g. if automation of physical tasks in manufacturing and logistics lags behind

Baumol's cost disease: non-automated jobs may experience rising costs, counteracting the overall economic impact

Gradual AI adoption/diffusion: technology may only be slowly integrated into various industries and adapted for different tasks, similar to the historical adoption of electricity.

Recap

Relevant actors include companies, nonprofits, governments, international organizations, and individuals.

Governance tools include disseminating information; standards, regulations, and laws; and financial incentives.

AI-driven automation might lead to explosive economic growth, if potential growth bottlenecks do not prevent this

Roadmap

1. Landscape: Which actors and tools can govern AI?
2. Growth: How will AI impact economic growth?
3. Distribution: How will costs, access to and power of AI be distributed?
4. Corporate governance: How can corporations be effectively governed?
5. National governance: How can national governments promote safety?
6. International governance: How can we govern AIs internationally?
7. Compute governance: Does governing compute help govern AIs?

Distribution

How will AI power, access, and costs and benefits distribute?

Distribution

Distribution of risks and benefits : will benefits and risks from AI be evenly or unevenly shared?

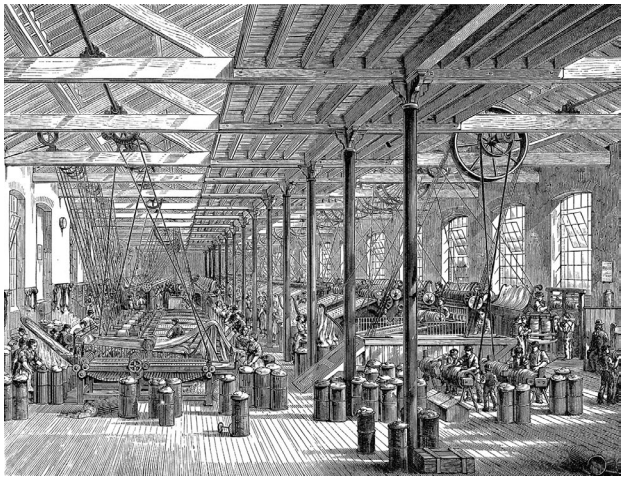
Distribution of access : will powerful AIs be kept in the hands of a small group of people, or broadly accessible to the public?

Distribution of power : will there only be a few highly powerful AIs, or many AIs with degrees of power?

Automation: short term

Jobs may be created in managing and checking AI systems throughout the economy. Overall productivity may also increased.

Previous technological revolutions like the industrial revolution ultimately led to more jobs and higher standard of living.



Automation: human-level AI

Human-level AI (HLAI) is by definition capable of doing every task that humans can do, at least as cheaply.

The development of HLAI would result in unprecedented mass unemployment, permanently. (Wages may go up before crashing.)

Humans may lose their main source of income, as well as the motivation and need to work and create.

This would create an unequal distribution of benefits, as people who own supercomputers would get the benefits of automating labor.

Levels of Access

Private AI models can only be used by a small set of people. Private companies and governments in particular may have restricted models, and these carry the risk of single points of failure and concentration of power/value lock-in.

Open weight AI models are freely available to the public, and can be edited and read. Such models carry the benefits and risks of wide access.

Structured access strikes a balance where the public can access AIs, with restrictions. Application programming interfaces (APIs) provide structured access. KYC could enable more intricate structured access.

Openness norms (1/2)

Traditionally, the norm in academia is for research to be openly shared.

However, this is not always the case. The Manhattan project was conducted in secrecy to avoid nuclear secrets from leaking to enemies.

While powerful AIs can be used for good, they can also be used with malicious intent, and thus total openness may not be responsible.

vs.

Openness norms (2/2)

Releasing model weights is a form of *irreversible* proliferation.

Unknown dangers: seemingly harmless developments in AI could be used in the development of dangerous AIs, or used in unforeseen ways.

However, interventions against openness have costs, like slowing down beneficial research.

Targeted and precise restrictions (structured access), as well as cost-benefit analysis of such restrictions, may be necessary.

Risks from open vs controlled models

Open	Controlled
Bottom-up malicious use: Rogue individuals and nonstate actors can pursue harmful activities, including creating biological, chemical, and physical weapons, cyberattacks on critical infrastructure, and more. Increased risk of tail event irreversible harms.	Top-down malicious use: where authorities or elites concentrate power. Totalitarian backsliding: use of AIs for surveillance, propaganda, and *unprecedented enforcement*.

Power: AI singleton

An **AI singleton** is the concept of an AI with decisive and lasting power over the world.

An AI singleton is made more likely if:

- AI companies are monopolies and face little competition.
- AI undergoes a fast takeoff where it surpasses all other intelligences and is able to suppress threats to its power.

Note that fast takeoff does not guarantee a singleton since simple structures can take down complex structures (e.g., viruses can kill humans destroy some AIs, bombs can destroy datacenters).

AI singleton: costs and benefits

A benefit of an AI singleton is that it would reduce competitive pressures and risks from collective action problems.

However, an AI singleton would be a single point of failure, potentially allowing a rogue AI or a malicious actor in control of it to pursue their goals unopposed.

Power: diverse AI ecosystem

A diverse ecosystem of AIs would have many capable AIs, none of which has decisive and lasting power over the others.

This may occur if many economic niches do not have increasing returns to intelligence/saturate, and thus there are many AIs of different capabilities.

Diverse AI ecosystem: benefits

Stability: diverse ecosystems are often more stable than ones with concentrated power.

- In agriculture, finance, and evolutionary biology, diversity has long term benefits.

Avoid single points of failure and malicious uses

- The failure of one AI will be mitigated by stability of AIs in charge of other processes.
- "Watchdog AIs" or similar concepts could counteract malicious AIs

Diverse AI ecosystem: costs

Multi-agent ecosystems have collective action problems and can lead to selfish traits.

Multiple AIs can exhibit collective emergent properties, creating feedback loops, and causing unanticipated failures.

Multiple AIs could collude to emulate the failure modes of a singleton.

Power among AIs may be distributed unequally (plausibly Pareto distributed), such that one or a few AIs are responsible for most of the impact on society.

Conclusions on distribution

AI power could distribute as a singleton, or a diverse ecosystem of AIs each with costs and benefits.

AI access could range from highly open to highly restricted. While openness entails dangers from the rogue individuals, restrictiveness entails top-down risks from elites.

The risks of AI are likely to be absorbed by the public unless governance measures are taken.

The benefits of AI are likely to accrue to Big Tech companies, but governance can create a more equitable distribution.

Roadmap

1. Landscape: Which actors and tools can govern AI?
2. Growth: How will AI impact economic growth?
3. Distribution: How will costs, access to and power of AI be distributed?
4. Corporate governance: How can corporations be effectively governed?
5. National governance: How can national governments promote safety?
6. International governance: How can we govern AIs internationally?
7. Compute governance: Does governing compute help govern AIs?

Corporate Governance

How can corporations be effectively governed?

Corporate governance

Multiple theories of what corporate governance should achieve.

Shareholder theory suggests that a corporation is obligated to maximize value for shareholders, or those with ownership in the company.

Stakeholder theory suggests that corporations should consider the interests of shareholders, as well as employees, customers, partners, and broader society.

We are concerned with how AI companies can steer AI development in a safe and beneficial direction.

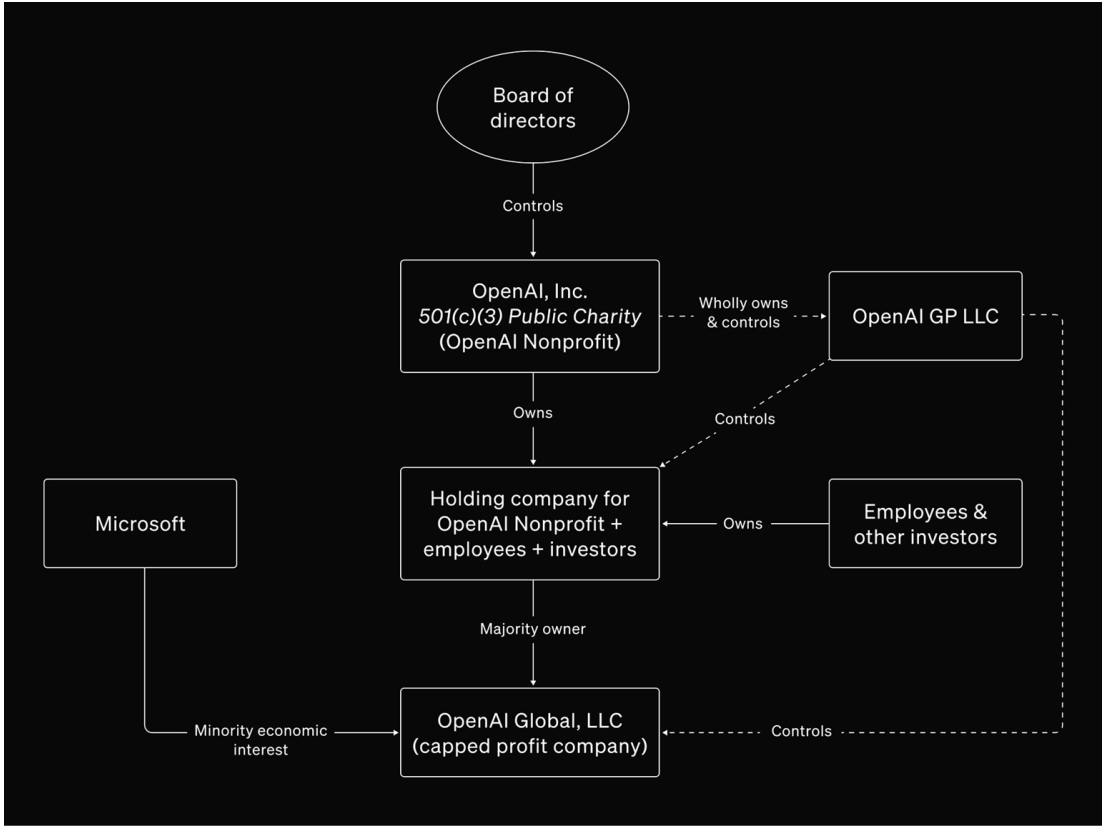
Legal structure (1/3)

AI companies can take different legal forms:

- C corporations or "C-corps" must maximize shareholder value: Google and Meta are C-corps
- Public benefit corporations (PBCs) can pursue public benefits: xAI and Anthropic are PBCs
- Limited partnerships (LPs), and limited liability companies (LLCs) are other common forms.

The place of incorporation can affect the rights and purposes of corporations. Many companies incorporate in Delaware because of its business friendly laws.

Legal structure (2/3)



Legal structure (3/3)

Statutes and bylaws allow companies to customize their official governance structures.

Companies must typically specify a mission statement upon incorporation.

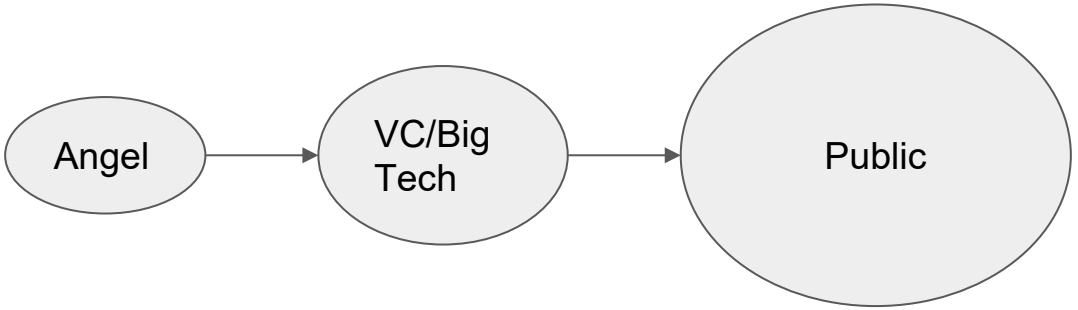
Statutes and bylaws can specify who has powers in crucial situations.

> Our mission is to ensure that artificial general intelligence—AI systems that are generally smarter than humans—benefits all of humanity.

Example: OpenAI's mission statement.

Ownership structure (1/3)

Companies are owned by shareholders, who elect a board of directors, who select executives, who directly manage the company.



Shareholders have the ability to vote on resolutions, voice their concerns to the company, and may file lawsuits against the company.

This is called **shareholder activism** .

Ownership structure (2/3)

The **board of directors** has **fiduciary duties** to act in the best interests of the company, which can vary depending on the mission. The board can set priorities, manages risk, selects senior management, and can form special committees.

Senior management (CEO, CTO, COO, etc.) directly runs the company. Boards may want to ensure that senior management cares about AI risk. Meta and xAI have CEOs with more decisive power; OpenAI and Anthropic are in-between; Google DeepMind does not.

Organizational units are internal teams like research, product, safety, and policy. AI companies could have special units for safety and societal considerations.

Ownership structure (3/3)

Ownership structures can often be customized.

Many companies issue two classes of shares: with and without voting rights.

Under OpenAI's "capped for profit" model, shareholders receive profits up to a certain threshold, and any further profits go to the OpenAI nonprofit entity.

In theory, there are many ways to govern a company. An AI company could even designate that under certain conditions, decision making be transferred to a committee representing broader society.

Conclusions on corporate governance

Corporate governance exists to balance the interests of shareholders and other stakeholders, including society.

Additionally, AI companies may need to consider novel governance structures, such as purpose trusts, capped-for-profit structures, and organizational units dedicated to societal risks.

Corporations generally have incentives to engage in regulatory capture and fight safety legislation.

National Governance

How can national governments promote safe AI?

National governments' role

Governments should enforce contracts, help solve collective action problems, make entities internalize their externalities, etc.

Tools available to national governments include:

1. Safety standards, regulation, and laws
2. Liability for harms from AI
3. Targeted taxation (e.g., value added tax on GPUs)
4. Public ownership of some AI systems

In addition, governments can take various measures aimed at:

- Improving societal resilience
- Developing AIs safely while not falling behind on innovation

Standards, Regulations, and Laws

Standards are formally written suggestions for best practices

- Standards can guide all aspects of the AI model life cycle.
- Standards are developed by governments, industry groups, nonprofits, and other organizations.

Regulations and Laws are legally binding requirements

- Regulations and laws are created by both legislative assemblies and regulatory agencies.
- Regulatory agencies and lawmakers are often influenced by special interest groups, and regulatory agencies are often under-resourced

Legal liability for AI harms

Legal liability determines who has to compensate for damages caused by AI, whether due to an accident or misuse.

Liability *internalizes risk* and promotes safety without regulations.

Limited tool: a company would not be able to cover all damages from a sufficiently massive catastrophe. A company probably cannot pay for the damages of an AI-enabled bioweapon.

AI developers have some liability due to background tort law.

Taxation and public ownership

Human labor is taxed (e.g., income tax), but automated labor is not. The government could fix this tax loophole.
Governments may **tax** AIs and compute to give money to people affected by mass automation

Public ownership of AIs can help governments ensure their safe development and redistribute profits, similar to how utilities are governed in many parts of the world.

Such policies are likely to be opposed by private companies.

Governments Can Improve Resilience

Resilience measures include cybersecurity, biosecurity, and "AI watchdogs" that monitor for dangerous AIs.

AIs can help improve cybersecurity by identifying weaknesses, but may also exploit them.

Governments can increase efforts to detect and counteract pathogens.

Certain AIs can monitor other AIs and identify threats.

Image generated with Stable Diffusion.

Not falling behind

If a country takes a safe and slow approach to AI development, they may be outpaced by the innovation of countries with looser guardrails.

An adversarial approach to AI policy may lead to international tensions, competitive dynamics, and unmanaged AI risks.

The US and China compete over technological development.

Not falling behind

Possible approaches to avoid falling behind while prioritizing safety include:

1. International cooperation (discussed later).
2. Export controls can stifle development in other countries.
3. Immigration policy can be used strategically to maintain a talent advantage.
4. Preventing model weights from being stolen by demanding better information security and strategic deception.
5. Intelligence can provide a view of the state of AI in other countries.

Conclusions on national governance

We discussed four areas of national governance of AIs:

1. Standards, regulations, and laws.
2. Legal liability for AI harms.
3. Improving resilience.
4. Not falling behind.

A combination of the approaches described in this chapter, and other approaches, should be taken for robust national AI governance.

Roadmap

1. Landscape: Which actors and tools can govern AI?
2. Growth: How will AI impact economic growth?
3. Distribution: How will costs, access to and power of AI be distributed?
4. Corporate governance: How can corporations be effectively governed?
5. National governance: How can national governments promote safety?
6. International governance: How can we govern AIs internationally?
7. Compute governance: Does governing compute help govern AIs?

International Governance

How can the international community govern AIs?

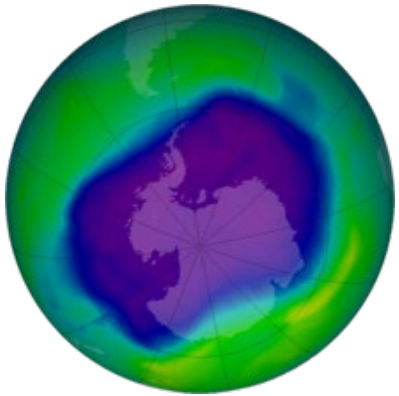
International governance

International governance can help share the benefits of AIs globally, and manage global risks.

International cooperation for the restriction of other hazardous technologies, like chemical (CWC), biological (BWC), and nuclear weapons (Atoms for Peace), has been difficult, but there are precedents.

Stages of international governance

Agenda setting: The community must recognize and agree that an issue requires international governance, and prioritize it.



Awareness of a depleting ozone layer prompted the Montreal Protocol.

Policymaking: International bodies can facilitate cooperation between countries to set laws and create formal guidelines.

The International Civil Aviation Organization sets aircraft safety standards.

Stages of international governance

Implementation and enforcement: Acting on agreements is difficult and requires resources, information, and coordination.

IAEA
International Atomic Energy Agency

The IAEA conducts inspections of nuclear facilities to ensure safety and compliance.

Evaluation, monitoring, and adjudication: International governance organizations, third party evaluators, or informal watchdogs may all play a part in reporting on compliance with guidelines.

The Organization for the Prohibition of Chemical weapons is one of several organizations that monitors chemical weapons.

International governance techniques (1/2)

Unilateral commitments: One actor (such as a country or company) makes a pledge regarding its use of technology. A variant is "*conditional leadership*": a leader will lead and will stop if people do not follow by a certain time. This helps resolve collective action problems.

- Compute commitments, commitments to work with AI safety institutes abroad

Norms and standards: Norms arise from discourse and precedent, standards are set by experts. Though nonbinding, they shape actions and align expectations.

- Scientific norms to publish intertwine US and Chinese AI research

International governance techniques (1/2)

Bilateral and multilateral talks: Talks build cooperation, and can help avoid technological races. They help rivals cooperate and negotiate.

- Carnegie Endowment for International Peace facilitates many track II talks between US and Chinese individuals

Summits and forums: Such events include a diverse set of stakeholders, can help set the agenda and create awareness for issues, and can serve as milestone moments.

- UK's AI Safety Summit, France's AI Action Summit

International governance techniques (2/2)

Governance organizations: International organizations are designated authority by members, and maintain governance efforts in the long term.

- UK AI Safety Institute, EU AI Office

Treaties: Strong form of governance with possible punishments for noncompliance, though they are difficult to negotiate.

- No current examples, and many of these treaties would not be trusted

Four relevant questions

Answering the following four questions can guide international cooperation:

1. When is the technology defense-dominant or offense-dominant?
2. Can compliance with international agreements be verified?
3. Is it catastrophic for international agreements to fail?
4. Can the production of the technology be controlled?

Is AI defense or offense dominant?

AI development would be *defense dominant* if good AIs could manage the threat of bad AIs.

AI development would be *offense dominant* if AIs cannot manage AIs.

Nuclear weapons are offense dominant because they are difficult to detect and easy to deploy.

AI seems likely to be offense dominant: it is easier to engineer pandemics using AI than counteract pandemics using AI.

Is AI compliance verifiable?

Detecting whether someone is developing or using an AI is not straightforward. They can do so discreetly.

We also don't have standard tests to verify if an AI model is safe, though we can work on developing standards.

Investing in strategies to verify the safety of AIs is a prerequisite to international regulation.

Compute security, as we'll discuss, could enable verification.

Is AI governance failure catastrophic?

Whether failure of compliance would be catastrophic determines how *permissible* international governance can be.

Example: Suppose there is an international agreement on the maximum amount of compute that can be used for an AI model.

If breaching the threshold is unlikely to be catastrophic, punishments can be after the fact and somewhat lenient.

If breaching the threshold may lead to catastrophic damages, enforcement must be proactive and harsh.

Can AI production be controlled?

Controlling access to factors of production—such as compute—can enable one or more responsible countries to regulate AIs.

Example: If the US had control over the production of compute, strategically controlling production may be easier than forming international agreements requiring many parties.

Regulating civilian AIs

Currently, much of the progress on AIs comes from the private sector.

Regulating civilian AI is difficult due to countries' resistance to international oversight and overlap between civilian and military use.

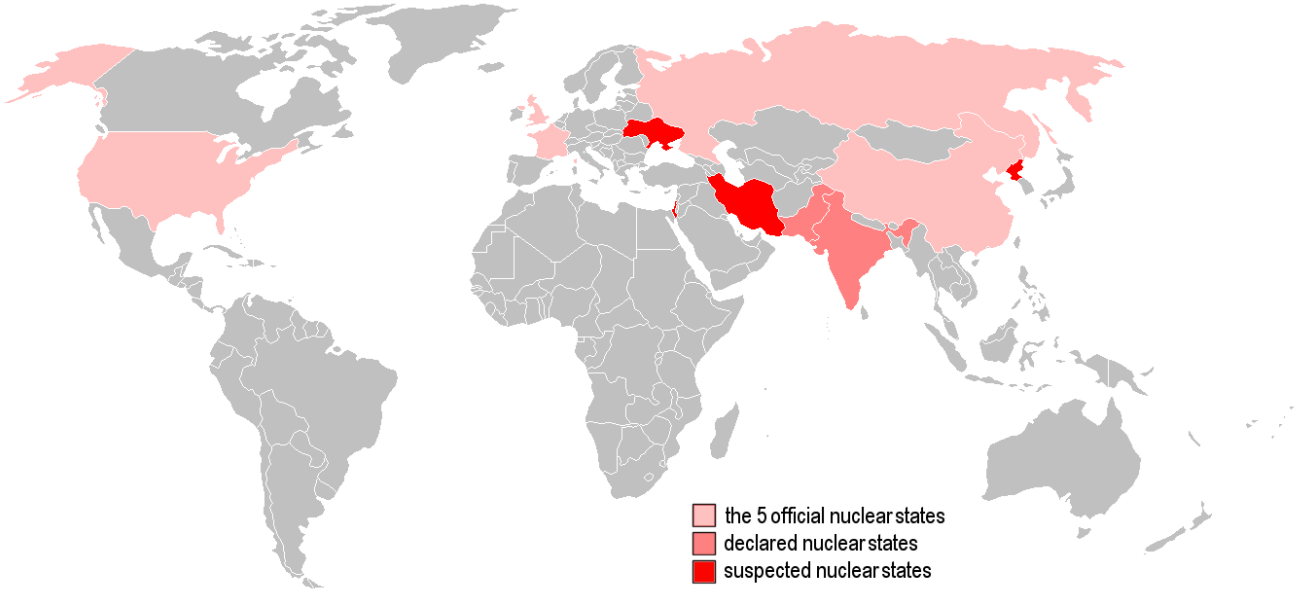
Certification, where international organizations assess whether jurisdictions have an acceptable regulatory regime, has been used for aviation and is a potential tool for governing AIs.

Regulating military AIs (1/2)

Non-prol t.

The Nucl owers
would nc would
share the

Non-prol fit
dramatic



Norms like "mutually assured destruction" may not be effective with AIs since AI usage is ubiquitous and dangers are unknown.

The Nuclear Non-Proliferation Treaty, while not signed by all nations or fully effective, set a precedent for global action on major risks.

Regulating military AIs (2/2)

Verification : Some countries may limit their development of AIs if they can verify that other countries are doing the same.

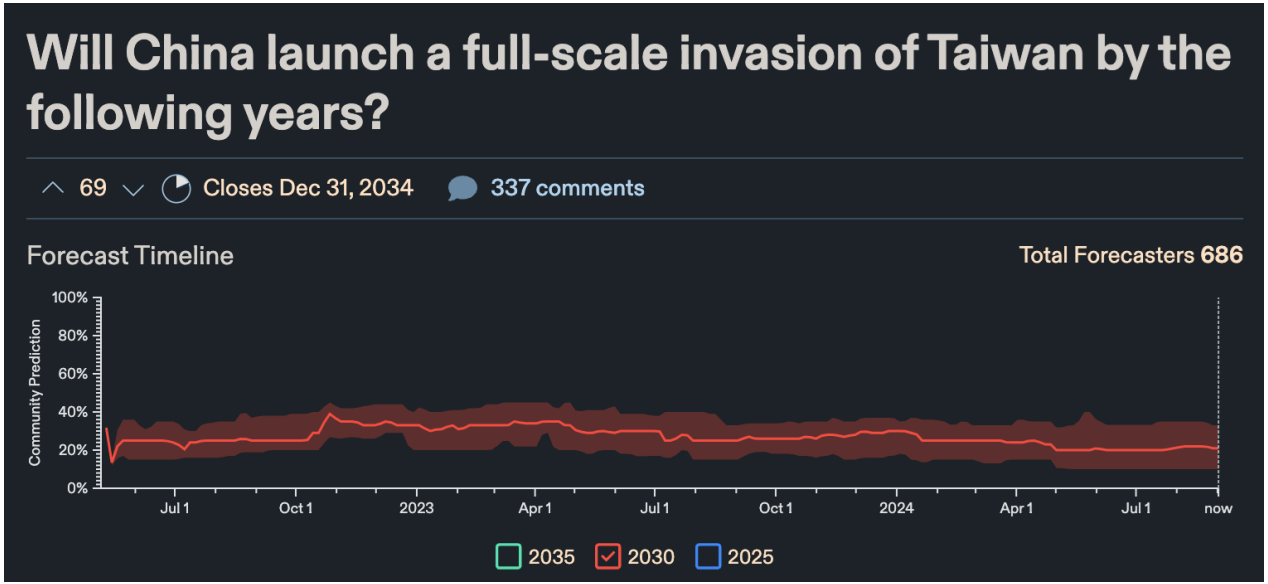
Verification requires inspections or compute security features.

Monopoly: AI monopolies could result from economies of scale or discontinuous/fast takeoff, or arise as the contracting of several influential AI entities.

Monopolies with sound corporate governance might be safer than complicated systems of norms, regulations, and enforcement.

Taiwan Invasion

A Chinese invasion of Taiwan would likely mean TSMC is not operational: this would cause a global depression, a multi-year AI slowdown, and possibly a great power conflict



Conclusions on international governance

International governance of AIs is necessary to distribute the benefits of AI and mitigate AIs global risks.

Stages: agenda setting, policymaking, implementation and enforcement, and evaluation, monitoring, and adjudication.

Tools: unilateral action, meetings, forums, international organizations, the creation of norms and standards, and binding treaties.

Features of AI determine the right governance measures.

Civilian and military AIs require unique governance approaches.

Compute Security and Governance

Does governing compute help govern AIs?

Compute or AI Chips

Compute is short for computation resources or computation power used for AIs.

1. Governing compute would help govern AIs.
2. Properties of compute make it governable.
3. Governing compute is more promising than governing other factors of AI.

Compute and AIs (1/2)

Compute is necessary for the development and deployment of AIs.

Currently, compute is of the form of computer chips like GPUs and TPUs, though chips used may change over time.

Compute is measured in floating point operations per sec (FLOP/s). A floating point operation is a simple mathematical operation like addition.

More compute → more parameters and train on larger datasets → better results.

Top AI models cost hundreds of millions to billions in compute, and this number is increasing.

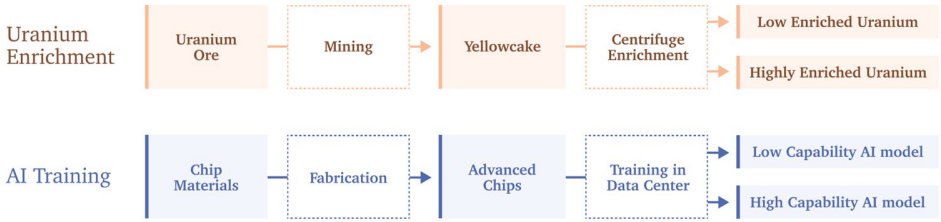
Compute and AIs (2/2)

AI research aims to reduce the compute requirements to develop advanced AIs, such as through the improvement of algorithms.

It's unclear if training a superintelligence will require, 10K, 100K, or 1mn GPUs. If 10K, then compute security will be less effective.

Data center

iPhone chip



Compute is governable

There are three primary factors of production for AIs: data, algorithms, and compute.

Compute is the most promising for governance because:

1. We can determine who has access to and is using compute.
2. We can enforce laws regarding compute.

Properties of compute

Physical: Unlike data or algorithms, compute is physical. Thus it is rivalrous (can only be used by one entity at a time) and trackable.

Quantifiable: We can know the quality and quantity of compute owned by anyone. Measurements like FLOP/s allow setting thresholds for regulation and monitoring compliance.

Excludable: There are only a few major semiconductor and chip firms that bodies like the U.S. and EU can regulate. AI companies are constantly upgrading their compute.

Compute Security

Compute Security is about adding low-level security features onto chips and is a research and engineering topic

1. Offline Licensing
2. Location Verification (with offline licensing, *geofencing* is possible)
3. Training Mode to Inference Mode switch (facilitates pause post-catastrophe or when automation becomes fast)
4. Training Compute Estimation

Compute security mechanisms can be privacy-preserving and many can be implemented in less than a year at a tamper-evident level

Conclusions

Compute is an indispensable part training and deploying AIs, and is more suited to be governed than data or algorithms.

Compute is physical, excludable, and quantifiable.

Governments have a number of tools to govern compute, including standards and regulations, and international trade controls.

Much of compute governance rests on compute security R&D

Recap

Corporate governance is a collection of regulatory tools that can be used to guide the direction of AI development by individual firms.

- Examples: public benefit corporation, capped profits

National governance can enable governments to control the development and deployment of AIs within their borders.

- Examples: standards and regulations, improving resilience

International governance brings together countries to create global agreements on how to create and use AIs.

- Examples: UN Committees, UK AI Safety Summit

Compute security may enable a verification regime and international coordination

Chapter recap

The landscape of AI governance includes relevant actors and tools.

AIs might "takeoff" in many different ways.

AIs' power, access, costs, and benefits can distribute in different ways.

Compute security is a promising way to govern AI.

Corporate, national, and international governance together govern AIs domestically and internationally, exerting pressure in different ways.