Summary of AI governance regulations

Mathis Embit

November 2023

Abstract

I wrote the following content for the Turing Seminar – An introduction to AGI Safety course of the MVA master. We had to choose a subject among a selection provided by our professors. I decided to write a "Summary of AI governance regulations". Work from my fellow classmates can be found on this LessWrong post.

Contents

0 Introduction

AI lacks meaningful binding safety standards despite its significant risks, as the rapid development and deployment of increasingly powerful systems outpace regulatory efforts. The unchecked pace has led to widespread harms such as mass disinformation, deep-fakes, and bias, posing threats to labor markets, political institutions, and national security. Urgent intervention by lawmakers is essential to establish and enforce safety standards, drawing parallels with existing regulatory practices for technologies like food, drugs, airplanes, and nuclear reactors to ensure AI benefits society without causing harm.

We will look at some recent research papers, notably "AI Governance Scorecard and Safety Standards Policy" from the Future of Life Institute (FLI) in which a comparison is made between the different AI governance proposals.

In order to simplify our discussion we try to locate each AI governance proposal in a conceptual graph of two axis: Regulation: no regulation → laws → pause → stop Cooperation: no cooperation → cooperation between the AI actors of the same country → cooperation within groups of countries (e.g. CERN) → international cooperation (no competition)

Let's first break down those regulation / cooperation axes and open discussion. After that we will briefly review some interesting strategies mentioned in the FLI paper.

1 How to put strategies into perspective?

1.1 How to regulate models?

How to regulate models (using FLI model requirements)? We can base our reflection on the regulation criterions proposed by FLI:

- Safety requirements even if not binding?
- Registration requirements?
- Third-party safety audit requirements?
- Burden of proof on developer to demonstrate safety?
- Quantitative risk bounds?
- Liability requirements?
- Compute limits?
- Doesn't exempt open source?
- Doesn't exempt LLMs?
- Doesn't exempt military AI?
- Calls for international regulatory body?
- Doesn't call for human replacement?

However some of them do not seem to be relevant because they are too specialized. I explain why I removed some criterions in the appendix A.

1.1.1 Fully open AI development

An unrestricted, fully open AI development strategy is seen as a catalyst for innovation and knowledge sharing. This approach encourages collaboration among diverse researchers and developers, leading to rapid progress. The transparency in open development helps identify and address issues like biases and ethical concerns, fostering a community-driven problem-solving approach. However, we can think of two major drawbacks to fully open development:

- Uncertainties regarding offense-defense balance.
- Reverse fine-tuning is very easy.

1.1.2 Pause or stop to AGI development

Potential consequences of pause frontier AI development:

- Economic impact: Slows down job creation and economic growth tied to AI industries, which can prevent countries from doing it.
- Global competition: Risks falling behind other countries in AI advancements, which can prevent countries from doing it.
- Delayed crucial problem solving: for example healthcare and climate change.
- Loss of talent: Skilled professionals may leave the field.

1.1.3 In between strategies

We pick some of the following criteria to construct the core of our custom AI governance strategy.

- **Model registration**

  The AI model registration process involves providing a detailed overview, including intended use, architecture, and dataset details. It documents hyperparameters, training processes, and performance metrics (accuracy, precision, recall, and F1 score).

- **3rd party audit**

  Auditing can occur at various levels, each addressing specific aspects of the development, deployment, and maintenance of artificial intelligence systems. These levels include:

  - Data Auditing: Focuses on the quality, source, and potential biases in the data used to train AI models.
  - Model Auditing: Examines the architecture, algorithms, and parameters of AI models.
  - Process Auditing: Evaluates the end-to-end development and deployment processes of AI systems.
  - Security Auditing: Assesses the security measures implemented to protect AI systems from vulnerabilities, attacks, and unauthorized access.
  - Performance Monitoring and Post-Deployment Auditing: Focuses on the ongoing monitoring of AI system performance after deployment. Addresses issues of system drift, identifies potential biases that may emerge in real-world usage, and ensures continuous improvement.

- **Safety demonstration by the developer**

  The responsibility for conducting a safety demonstration lies with the developer as he possesses an in-depth understanding of the technical intricacies of the AI system, including its architecture, algorithms, and potential vulnerabilities. Here are some common elements and types of demonstrations that developers might consider:

  - Explanatory Walkthroughs
  - Bias Analysis Showcase

  - Robustness Testing Scenarios
  - Privacy Protection Mechanisms
  - Human Oversight Integration
  - Fail-Safe Mechanisms in Action
  - Continuous Monitoring

- **Risk bounds**

  We may define risk bounds over some measurements such as:

  - Bias Measurement: Use metrics such as disparate impact, equalized odds, or statistical parity to quantitatively measure bias in AI models. Calculate the difference in model performance across different demographic groups and assess fairness using appropriate fairness metrics.
  - Accuracy Measurement: Calculate standard accuracy metrics, such as precision, recall, F1 score, and confusion matrices, to measure the performance of the AI model.
  - Robustness Testing: Perform robustness testing by subjecting the AI system to adversarial inputs or edge cases. Measure the model's performance under these conditions and quantify robustness using metrics like adversarial accuracy or sensitivity to input variations.

- **Defining responsibility agreement for the developer and the user**

  Examples of existing agreements of this type. We can get inspired by examples coming from other fields. Concerning individual use we can think of:

  - Car industry is highly regulated, and the user needs to get a driving license and respect some well defined laws.
  - Firearms...

- **Compute limits**

  Current compute power: https://openai.com/research/ai-and-compute

  Prediction for future compute power: https://en.wikipedia.org/wiki/Moore's_law

  Compute governance addresses 3 key issues:

  - Management of physical space and energy demand.
  - Can be used to improve AI safety by restricting compute access to non-safety-aligned actors.
  - A trend tends to reverse: computer science research costs which became low, requires an important infrastructure budget again.

- **Allowing military free use?**

  Military regulation example, in the US:

  https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy/

1.2 Cooperation

Once criterions has been chosen we need to pack all this into an institution. The choice of this institution is also a key aspect of the governance strategy. The choice of the institution depends on the choice we want to make on the cooperation axis.

1.2.1 Non-governmental actors initiative

Big tech companies propose a view on AI safety. Making the leading experiments they need to be seen as trustful. For example Deepmind proposed a framework for sociotechnical AI safety evaluation:

- Layer 1: Capability: AI systems and their technical components being routinely evaluated in isolation.
- Layer 2: Human interaction: Human-centered testing can shed light on potential externalities created by specific use cases or applications of AI systems.
- Layer 3: Systemic impact: Detecting effects of an AI system on the broader systems, such as society, the economy, and the natural environment, effects that may only emerge as the AI system is deployed at large scale.

They stress on the model evaluation, being the main tools we have for AI risk assessment. They provide a blueprint for how to guard against extreme risks while developing and deploying a model, with evaluation embedded throughout. They initiate some guidelines but they do not have the power to make them obligatory for other companies. Doing so guides their own research but does not stop the race to AGI for example.

AI evaluation companies such as Arc evals also propose safety standards for AI governance strategies. For example ARC evals proposed Responsible Scaling Policies (RSP) which specifies what level of AI capabilities an AI developer is prepared to handle safely with their current protective measures. Their goal is to maintain AI research in a safe region, balancing between protective measures and dangerous capabilities. Solutions like ARC evals would provide a useful help to authorities to audit AI systems development. That goes in the sense of GovAI which writes: “Governments can advance the development of standards by working with stakeholders to create a robust ecosystem of safety testing capability and auditing organizations, seeding a third-party assurance ecosystem”.

In Frontier AI Regulation: Managing Emerging Risks to Public Safety GovAI claims that “self-regulation is unlikely to provide sufficient protection against the risks from frontier AI models: government intervention will be needed. Without government intervention, there would be a lack of compliance”. Let’s see what governments have done to this day.

1.2.2 Government initiative

The EU is a pioneer in regulation. They recently proposed the EU AI Act Compromise Proposal.

In the paper Managing AI Risks in an Era of Rapid Progress researchers propose reorienting technical R&D to AI safety and want national institutions and an international governance framework. However, reorienting technical R&D to AI safety would indeed require international cooperation. If a government decides to reorient AI research to mainly safety, other governments or companies of other governments will take advantage of it to take the lead in the race to AGI. We can clearly see that, even in Biden’s executive order: “The Executive Order establishes new standards for AI safety and security, protects Americans’ privacy, advances equity and civil rights, stands up for consumers and workers, promotes innovation and competition, advances American leadership around the world, and more”. We can legitimately doubt international cooperation.

2 Review of some strategies

2.1 GovAI’s International AI Organization (IAIO)

Model registration	3rd party audit	Safety demonstration by the developer	Risk bounds	Responsibility agreement	Compute limits	Allowing military free use?
Yes	Yes	No	No	Yes	No	No

In "Frontier AI Regulation: Managing Emerging Risks to Public Safety", GovAI proposes three building blocks for the regulation of frontier models:

1. standard-setting processes to identify appropriate requirements for frontier AI developers
2. registration and reporting requirements to provide regulators with visibility into frontier AI development processes
3. mechanisms to ensure compliance with safety standards for the development and deployment of frontier AI models.

GovAI researchers think that industry self-regulation is an important first step but a government intervention will be needed to create standards and to ensure compliance with them. GovAI coordinates in the regulation/cooperation graph would be:

- Regulation = soft regulation (requires legal procedures but no hard bounds such as compute limits)
- Cooperation = international

Note: as seen in the FLI paper hard bounds (proof by developer, quantitative risk bounds, compute limits) tends to appear when a first layer of legal issues is treated (registration, audit, ...). Therefore the regulation axis would look like: no regulation → soft regulation → hard regulation bounds → pause → stop

2.2 UK Government "Emerging Processes [...] Safety"

Model registration	3rd party audit	Safety demonstration by the developer	Risk bounds	Responsibility agreement	Compute limits	Allowing military free use?
No	No	No	No	No	No	No

The UK government clearly supports a pro-innovation approach. The first words of their paper are: "The UK recognises the enormous opportunities that AI can unlock across our economy and our society. However, without appropriate guardrails, such technologies can pose significant risks." They do not insist on cooperation but still encourage the construction of government and international governance institutions to get involved: "the appropriate government and international governance institutions are still being considered and we recognise that limits the ability of frontier AI organizations to share information with governments, even where it would be desirable". UK government coordinates in the regulation/cooperation graph would be:

- Regulation = almost none
- Cooperation = desirable

2.3 Anthropic's "Responsible Scaling Policy" (RSP)

Model registration	3rd party audit	Safety demonstration by the developer	Risk bounds	Responsibility agreement	Compute limits	Allowing military free use?
No	No	No	No	No	No	No

Anthropic is regulating itself, probably hoping it inspires others to do so. Anthropic coordinates in the regulation/cooperation graph would be:

- Regulation = almost none but internally follows safe research protocols
- Cooperation = no mention since it's about self regulation

2.4 "Managing AI Risks in an Era of Rapid Progress"

Model registration	3rd party audit	Safety demonstration by the developer	Risk bounds	Responsibility agreement	Compute limits	Allowing military free use?
Yes	Yes	No	No	Yes	No	No

Talking about the governance measures, the paper clearly says: "We urgently need national institutions and international governance to enforce standards to prevent recklessness and misuse". The paper calls for high regulation: "Regulators should require registration of frontier systems in development, whistleblower protections, incident reporting, and monitoring of model development and supercomputer usage. Regulators also need access to advanced AI systems before deployment to evaluate them for dangerous capabilities such as autonomous self-replication, breaking into computer systems, or making pandemic pathogens widely accessible." "Managing AI Risks in an Era of Rapid Progress" consensus coordinates in the regulation/cooperation graph would be:

- Regulation = soft
- Cooperation = international

2.5 PauseAI Proposal

Model registration	3rd party audit	Safety demonstration by the developer	Risk bounds	Responsibility agreement	Compute limits	Allowing military free use?
Yes	Yes	Yes	No	Yes	Yes	No

PauseAI proposal requires strong international cooperation as it claims that: "Individual countries can and should implement this measure right now". As the proposal's name says, the regulation is high and calls for hard bounds such as: "Implement a temporary pause on the training of AI systems more powerful than GPT-4". PauseAI Proposal coordinates in the regulation/cooperation graph would be:

- Regulation = strong (including hard bounds such as compute limits)
- Cooperation = international

2.6 FLI proposal

Model registration	3rd party audit	Safety demonstration by the developer	Risk bounds	Responsibility agreement	Compute limits	Allowing military free use?
Yes	Yes	Yes	Yes	Yes	No	No

"AI Governance Scorecard and Safety Standards Policy" by FLI makes a comparison and proposes a framework of hard requirements before training and deploying. The paper also makes a distinction between commercial AI which is globally good for humanity and AGI pursuit which won't benefit anyone. What they want is:

- not pause
- stop training ever-larger models until they meet reasonable safety standards

Their framework is opposed to the main AI companies strategies which rely on corporate self-regulation or voluntary commitments. The framework consists of 4 safety levels. Each level is defined by specific technical criteria. Each level calls for hard requirements before training and deploying, enforced by national or international governing bodies. FLI coordinates in the regulation/cooperation graph would be:

- Regulation = strong but does not limit compute
- Cooperation = preferably international

3 Conclusion

As a conclusion, let's draw the regulation / cooperation graph and put the few strategies we looked at, leaving space for imagination to construct new strategies, or new axes worth using!

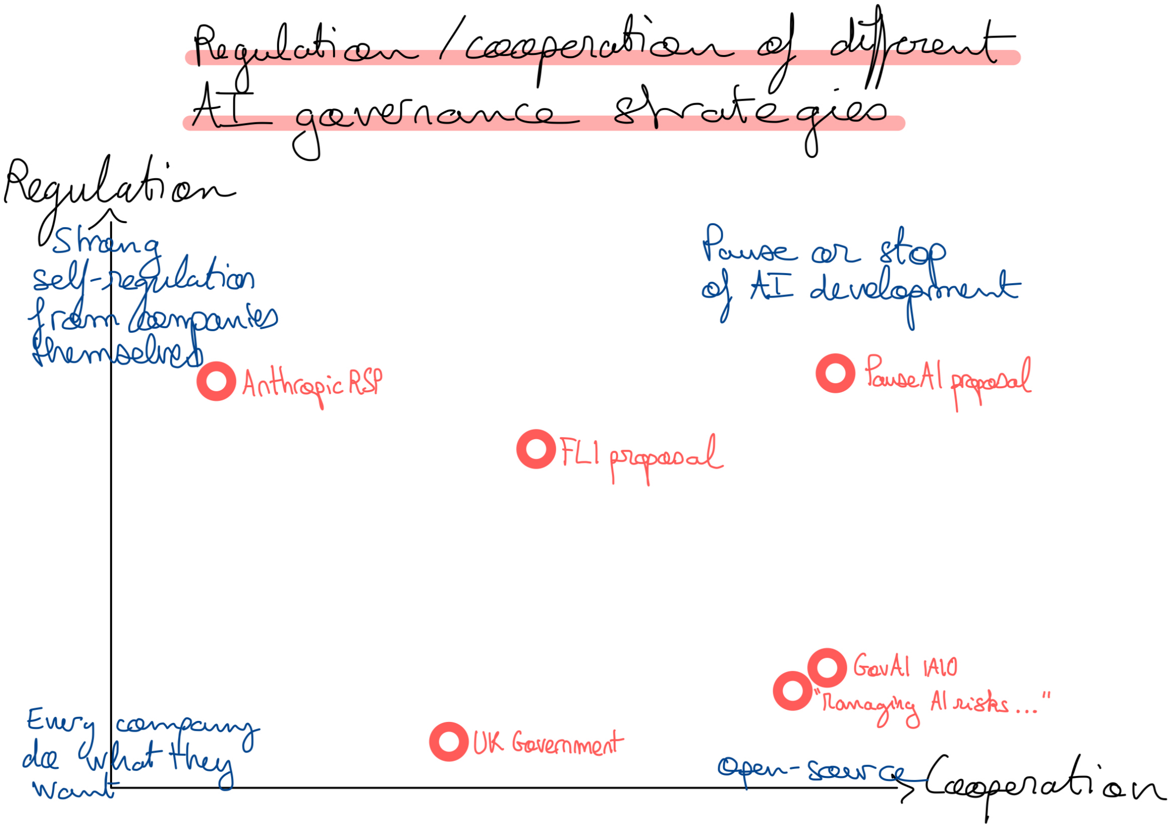

Figure 1: Attempt to place the AI governance strategies on the previously defined regulation and cooperation axes.

A Appendix

The criterions I decided to remove are:

- Exemption concerning open-source. This criterion lacks generality. Open-source is one way of developing AI among many others. That's why I decided to define a cursor on a regulation spectrum where open-source is one of the two extremes.
- Exemption in regulation concerning LLMs. LLM may currently be the most impactful model but other architectures may surpass it in the future. This criterion lacks generality, particularly in time.

- Calls for international regulatory body. This call also lacks generality. A strategy may want to call for a group of country regulatory body. That's why I decided to define a cursor on a cooperation spectrum where "international regulatory body call" is a consequence of a strategy crossing a threshold toward international cooperation.
- Doesn't call for human replacement? I think it won't be exaggerated to say that the majority of humans don't want to be replaced. We can see that by the fact that only one strategy of the FLI paper calls for human replacement.

Though, it may be wise to let the "military free use" criteria. Naively, we observe that the defense sector sometimes helps to develop helpful technologies such as Arpanet but also tends to harm, for example by the development of the atomic bomb.

References

[1] AI Governance Scorecard and Safety Standards Policy, Future of Life Institute.
https://futureoflife.org/wp-content/uploads/2023/11/FLI_Governance_Scorecard_and_Framework.pdf

[2] Frontier AI Regulation: Managing Emerging Risks to Public Safety.
https://arxiv.org/abs/2307.03718

[3] Managing AI Risks in an Era of Rapid Progress.
https://managing-ai-risks.com/

[4] Sociotechnical Safety Evaluation of Generative AI Systems.
https://arxiv.org/abs/2310.11986

[5] Model evaluation for extreme risks, Deepmind.
https://browse.arxiv.org/pdf/2305.15324.pdf

[6] Responsible Scaling Policies, Model Evaluation and Threat Research.
https://metr.org/blog/2023-09-26-rsp/

[7] President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence.
https://www.lesswrong.com/posts/g5XLHKyApAFXi3fso/president-biden-issues-executive-order-on-safe-secure-and