AI Governance Scorecard and Safety Standards Policy

Evaluating proposals for AI governance and providing a regulatory framework for robust safety standards, measures and oversight.

30th October 2023

Last update: 6th November 2023

policy@futureoflife.org

Introduction

AI remains the only powerful technology lacking meaningful binding safety standards. This is not for lack of risks. The rapid development and deployment of ever-more powerful systems is now absorbing more investment than that of any other models. Along with great benefits and promise, we are already witnessing widespread harms such as mass disinformation, deepfakes and bias - all on track to worsen at the currently unchecked, unregulated and frantic pace of development. As AI systems get more sophisticated, they could further destabilize labor markets and political institutions, and continue to concentrate enormous power in the hands of a small number of unelected corporations. They could threaten national security by facilitating the inexpensive development of chemical, biological, and cyber weapons by non-state groups. And they could pursue goals, either human- or self-assigned, in ways that place negligible value on human rights, human safety, or, in the most harrowing scenarios, human existence.

Despite acknowledging these risks, AI companies have been unwilling or unable to slow down. There is an urgent need for lawmakers to step in to protect people, safeguard innovation, and help ensure that AI is developed and deployed for the benefit of everyone. This is common practice with other technologies. Requiring tech companies to demonstrate compliance with safety standards enforced by e.g. the FDA, FAA or NRC keeps food, drugs, airplanes and nuclear reactors safe, and ensures sustainable innovation. Society can enjoy these technologies' benefits while avoiding their harms. Why wouldn't we want the same with AI?

With this in mind, the Future of Life Institute (FLI) has undertaken a comparison of AI governance proposals, and put forward a safety framework which looks to combine effective regulatory measures with specific safety standards.

AI Governance Scorecard

Recent months have seen a wide range of AI governance proposals. FLI has analyzed the different proposals side-by-side, evaluating them in terms of the different measures required. The results can be found below. The comparison demonstrates key differences between proposals, but, just as importantly, the consensus around necessary safety requirements. The scorecard focuses particularly on concrete and enforceable requirements, because strong competitive pressures suggest that voluntary guidelines will be insufficient.

The policies fall into two main categories: those with binding safety standards (akin to the situation in e.g. the food, biotech, aviation, automotive and nuclear industries) and those without (focusing on industry self-regulations or voluntary guidelines). For example, Anthropic's Responsible Scaling Policy (RSP) and FLI's Safety Standards Policy (SSP) are directly comparable in that they both build on four AI Safety Levels – but where FLI advocates for an immediate pause on AI not currently meeting the safety standards below, Anthropic's RSP allows development to continue as long as companies consider it safe. The FLI SSP is seen to check many of the same boxes as various competing proposals that insist on binding standards, and can thus be viewed as a more detailed and specific variant alongside Anthropic's RSP.

Table 1: A summary of the AI governance playing field going in to the November 1-2 UK AI Summit

		Binding regulation proposed						Exemptions				
AI governance proposal	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?
	Proposes specific safety measures for AI systems, even if compliance is not enforceable.	Mandates the recording and submission of specific details about an AI system prior to its training and deployment.	Stipulates that AI systems must undergo a systematic and independent examination to ensure safety measures are met.	Obliges AI developers to proactively provide evidence or justification of the safety of their systems prior to training and deployment.	Defines numerical thresholds or limits pertaining to the potential risks or harm an AI system might pose.	Outlines the responsibilities and legal consequences for developers or users should their AI system cause harm or operate outside of its defined parameters.	Sets boundaries on the computational resources or power that an AI system can use.	Does not exempt open-source or widely released AI models from the requirements of the proposal.	Does not exempt large language models from the requirements of the proposal.	Does not exempt military training and deployment of AI systems from the requirements of the proposal.	Advocates for the establishment of a global organization responsible for overseeing AI safety and standards.	Refrains from advocating for the replacement of humanity with autonomous AI.
Sutton's "AI Succession"	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Andreessen's "Techno-Optimist Manifesto"	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
PAI's "Deployment Guidance"	✓	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✓
Anthropic's "Responsible Scaling Policy" (RSP)	✓	✗	✗	✗	✗	✗	✗	✓	✓	✓	✗	✓
UK Government "Emerging Processes...Safety"	✓	✗	✗	✗	✗	✗	✗	✓	✓	✓	✗	✓
President Biden's Executive Order	✓	✓	✗	✗	✗	✗	✗	✓	✓	✓	✗	✓
GovAI's International AI Organization (IAIO)	✓	✓	✓	✗	✗	✓	✗	✓	✓	✓	✓	✓
French EU AI Act revision proposal	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✓	✓
EU AI Act Compromise Proposal	✓	✓	✓	✓	✗	✗	✗	✓	✓	✗	✓	✓
Chinese AI Policy	✓	✓	✗	✗	✗	✓	✗	✓	✓	✓	✗	✓
AI Treaty Open Letter	✓	✗	✗	✗	✓	✗	✓	✓	✓	✓	✓	✓
Treaty on AI Safety and Collaboration (TAISC)	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓
"Managing AI Risks in an Era of Rapid Progress"	✓	✓	✓	✗	✗	✓	✗	✓	✓	✓	✓	✓
Ditchley Declaration	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓	✓
FLI "Safety Standards Policy" (SSP)	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓
PauseAI's proposal	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓
Yudkowsky's "Shut it All Down"	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Non-AI Examples												
FDA (Food & Drug Administration)	✓	✓	✓	✓	✓	✓						
FAA (Federal Aviation Administration)	✓	✓	✓	✓	✓	✓						
NRC (Nuclear Regulatory Commission	✓	✓	✓	✓	✓	✓						

Please report comments or corrections to contact@futureoflife.org

FLI Safety Standards Policy (SSP)

Taking this evaluation and our own previous policy recommendations into account, FLI has outlined an AI safety framework that incorporates the necessary standards, oversight and enforcement to mitigate risks, prevent harms, and safeguard innovation. It seeks to combine the "hard-law" regulatory measures necessary to ensure compliance – and therefore safety – with the technical criteria necessary for practical, real-world implementation.

The framework contains specific technical criteria to distinguish different safety levels. Each of these calls for a specific set of hard requirements before training and deploying such systems, enforced by national or international governing bodies. While these are being enacted, FLI advocates for an immediate pause on all AI systems that do not meet the outlined safety standards.

Crucially, this framework differs from those put forward by AI companies (such as Anthropic's 'Responsible Scaling Policy' proposal) as well as those organized by other bodies such as the Partnership on AI and the UK Task Force, by calling for legally binding requirements – as opposed to relying on corporate self-regulation or voluntary commitments.

The framework is by no means exhaustive, and will require more specification. After all, the project of AI governance is complex and perennial. Nonetheless, implementing this framework, which largely reflects a broader consensus among AI policy experts, will serve as a strong foundation.

Table 2: FLI's Proposed Policy Framework

Classification	Hardware trigger	Capabilities trigger *(1-3 based on Anthropic's classification)*	Requirements for training	Requirements for deployment
ASL-1	None	**Negligible potential for harm:** Systems which pose no meaningful catastrophic risk, for example a 2018 LLM or an AI system that only plays chess.	None	None
ASL-2	None	**Potential for minor harm:** Systems that show early signs of dangerous capabilities – for example ability to give instructions on how to build bioweapons – but where the information is not yet useful due to insufficient reliability or not providing information that e.g. a search engine couldn't.	Registration with national authority with juristiction over the lab	**Safety audits** by national authorities wherever the model can be used, including blackbox and whitebox red-teaming
ASL-3	100 yotta-FLOP (10^{26})	**Potential for major harm:** Systems that substantially increase the risk of catastrophic misuse compared to non-AI baselines (e.g. search engines or textbooks) OR that show low-level autonomous capabilities, alone or in combination with other available techniques OR that classify as "very capable foundation models" under provisions of the EU AI act.	Pre-approval of safety plan by national authority with juristiction over the lab	**Quantitative safety bounds:** National authorities wherever the model can be used must approve lab-submitted assesment bounding risk of major harm below authorized levels.
ASL-4	ronnaFLOP (10^{27})	**AGI potential:** Systems with potential to enable AGI alone or in combination with other available techniques, where AGI is defined as AI capable of performing all economically valuable cognititive tasks at human expert level.	Pre-approval of safety plan by IAIA	**Provable safety:** The IAIA must certify lab-submitted formal verification that the model provably meets required specifications, including cybersecurity, controllability, a non-removable kill-switch, alignment with human values, and robustness to malicious use.

Clarifications

Triggers: A given ASL-classification is triggered if either the hardware trigger or the capabilities trigger applies.

Registration: This includes both training plans (data, model and compute specifications) and subsequent incident reporting. National authorities decide what information to share.

Safety audits: This includes both cybersecurity (preventing unauthorized model access) and model safety, using whitebox and blackbox evaluations (with/without access to system internals).

Responsibility: Safety approvals are broadly modeled on the FDA approach, where the onus is on AI labs to demonstrate to government-appointed experts that they meet the safety requirements.

IAIA international coordination: Once key players have national AI regulatory bodies, they should aim to coordinate and harmonize regulation via an international regulatory body, which could be modeled on the IAEA – above this is referred to as the IAIA ("International AI Agency") without making assumptions about its actual name. In the interim before the IAIA is constituted, ASL-4 systems require UN Security Council approval.

Liability: Developers of systems above ASL-1 are liable for harm to which their models or derivatives contribute, either directly or indirectly (via e.g. API use, open-sourcing, weight leaks or weight hacks).

Kill-switches: Systems above ASL-3 need to include non-removable kill-switches that allow appropriate authorities to safely terminate them and any copies.

Risk quantification: Quantitative risk bounds are broadly modeled on the practice in e.g. aircraft safety, nuclear safety and medicine safety, with quantitative analysis producing probabilities for various harms occurring. A security mindset is adopted, whereby the probability of harm factors in the possibility of adversarial attacks.

Compute triggers: These can be updated by the IAIA, e.g. lowered in response to algorithmic improvements.

Why regulate now?

Until recently, most AI experts expected truly transformative AI impact to be at least decades away, and viewed associated risks as "long-term". However, recent AI breakthroughs have dramatically shortened timelines, making it necessary to consider these risks now. The plot below (courtesy of the Metaculus prediction site) shows that the number of years remaining until (their definition of) Artificial General Intelligence (AGI) is reached has plummeted from

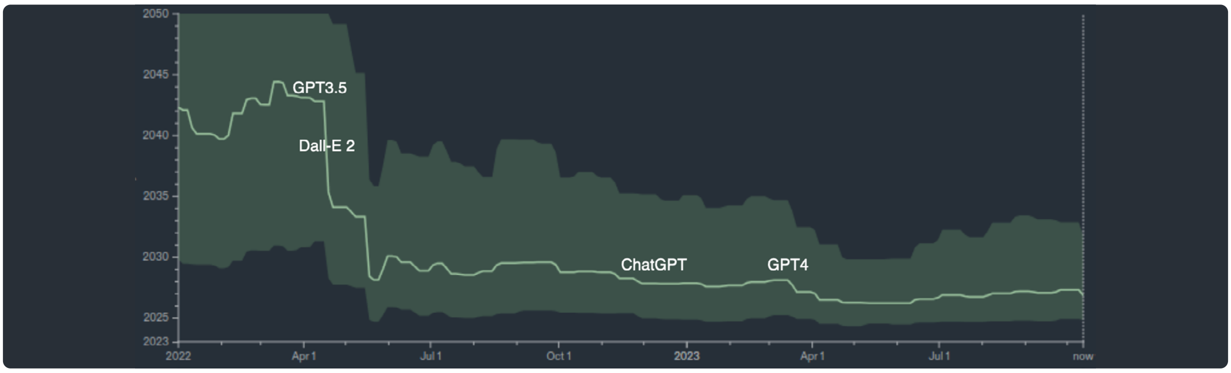

Image: 'When will the first weakly general AI system be devised, tested, and publicly announced?' at Metaculus.com twenty years to three in the last eighteen months, and many leading experts concur. For example, Anthropic CEO Dario Amodei predicted AGI in 2-3 years, with 10-25% chance of an ultimately catastrophic outcome. AGI risks range from exacerbating all the aforementioned immediate threats, to major human disempowerment and even extinction – an extreme outcome warned about by industry leaders (e.g. the CEOs of OpenAI, Google DeepMind & Anthropic), academic AI pioneers (e.g. Geoffrey Hinton & Yoshua Bengio) and leading policymakers (e.g. European Commission President Ursula von der Leyen and UK Prime Minister Rishi Sunak).

Reducing risks while reaping rewards

Returning to our comparison of AI governance proposals, our analysis revealed a clear split between those that do, and those that don't, consider AGI-related risk. To see this more clearly, it is convenient to split AI development crudely into two categories: commercial AI and AGI pursuit. By commercial AI, we mean all uses of AI that are currently commercially valuable (e.g. improved medical diagnostics, self-driving cars, industrial robots, art generation and productivity-boosting large language models), be they for-profit or open-source. By AGI pursuit, we mean the quest to build AGI and ultimately superintelligence that could render humans economically obsolete. Although building such systems is the stated goal of OpenAI, Google DeepMind, and Anthropic, the CEOs of all three companies have acknowledged the grave associated risks and the need to proceed with caution.

The AI benefits that most people are excited about come from commercial AI, and don't require AGI pursuit. AGI pursuit is covered by ASL-4 in the FLI SSP, and motivates the compute limits in many proposals: the common theme is for society to enjoy the benefits of commercial AI without recklessly rushing to build more and more powerful systems in a manner that carries significant risk for little immediate gain. In other words, we can have our cake and eat it too. We can have a long and amazing future with this remarkable technology. So let's not pause AI. Instead, let's stop training ever-larger models until they meet reasonable safety standards.

About the Future of Life Institute

The Future of Life Institute (FLI) is an independent non-profit organization that works to steer transformative technologies to benefit humanity and avoid catastrophic risks. Back in 2017, FLI organized a conference in Asilomar, California to formulate one of the earliest artificial intelligence (AI) governance instruments: the "Asilomar AI principles." The organization has since become one of the leading voices on AI policy in Washington D.C. and Brussels, and is now the civil society champion for AI recommendations in the United Nations Secretary General's Digital Cooperation Roadmap. In March, FLI – joined by over 30,000 leading AI researchers, professors, CEOs, engineers, and others – called for a pause of at least six months on the largest and riskiest AI experiments, to allow time for binding safety standards (such as those discussed above) to be implemented. The letter sparked United States Senate hearings, a formal reply from the European Parliament, and a call from UNESCO to implement a global ethical framework for AI.

FLI is grateful to Dr. Peter Park from MIT for his invaluable research underlying the scorecard.