A New Regulator for Advanced AI Systems

Why Action Is Needed – A Statement on AI's Risk of Human Extinction:

CEOs of America's leading AI companies, Nobel Prize winners, and national security experts have all warned that AI development could cause human extinction, and that mitigating this risk should be a global priority.

Specialized AIs can boost our economy, protect our warfighters, drive better and simpler public services, and enable innovation. Superintelligent AIs threaten our national security.

The American people must stay in control. To protect our nation, and secure AI's benefits for our future, the United States must establish and maintain control over the development of dangerous AI systems.

The Solution – A Framework to Protect the American People from the Risks of Advanced AI Systems

In November 2023, the US government established the US AI Safety Institute (AISI) to advance our understanding of advanced AI and protect the American people from frontier AI threats. We propose **passing permanent statutory authorization for a renamed AISI, the US AI *Security* Institute, as an independent AI security and regulatory agency** outside the Department of Commerce, with the authority to regulate, oversee, and enforce security and safety standards for frontier AI models[1] This regulator would ensure that companies developing AI models above certain compute thresholds and general intelligence benchmarks comply with rigorous security and safety protocols. This would allow the US to harness the benefits of practical AI while mitigating risks posed by the uncontrolled development of superintelligent AI.

The regulator would be led, as is the case for many regulators in the US, by a multi-member panel of regulatory commissioners with staggered terms of office, nominated by the President and confirmed by the Senate. The membership would be roughly equal between the two parties, with no more than half-plus-one of the members from each party. The President would designate one commissioner as Chairman. (This model is similar to, e.g., the US Securities Exchange Commission).

---

[1] Note: there are a wide range of costs and benefits to moving AISI out of the Department of Commerce. Congress might feasibly choose instead to retain it in its current location, locate it within another government Department (e.g., Energy, Homeland Security), or create a CFIUS-like coordinating body.

There are meaningful legal considerations (e.g., US Persons rules) that mean that AISI would have difficulty interfacing with US AI companies if it was located in the Department of Defense or the Intelligence Community, so we specifically recommend against those options.

To align incentives and to ensure prompt regulatory action, AISI could be funded through fees from those regulated by it, similar to the FDA's proven approach to industry partnership.

Key aspects of the AI regulator mandate would include:

- Licensing frontier AI developers to ensure AI models do not pose risks to the American public or national security before, during, and after development.
- Prohibiting dangerous AI capabilities, such as unauthorised replication, environmental breakout, and autonomous self-improvement.
- Oversight of high-computation AI models and applications that present catastrophic or extinction level risks.
- Establishing security and product safety standards for the design, development, deployment, and monitoring of AI systems.
- Threat, scenario, and trend assessment of security, product safety, and risk impacts of AI.

The Licensing Framework

At the core of the regulator's power is a **three-tiered licensing system** aimed at managing the development and deployment of frontier AI models above critical compute thresholds. These licenses ensure that only AI developers and operators that meet safety requirements can proceed with their work.

1. **Training License**
   For AI developers aiming to train models that exceed a set computational power threshold, set at 10^25 FLOP. Inspired by other current industrial security and industrial safety boards, regulated applicants must present detailed risk mitigation plans for managing developing and deploying AI for their intended use, including shutdown procedures for AI systems that pose unacceptable levels of risk.
2. **Compute License**
   Required for cloud service providers and data centres operating above 10^17 FLOP/s. The compute license ensures that large-scale computational power is not misused for unregulated AI development. Licensees must implement hardware tracking and know-your-customer (KYC) requirements to maintain transparency and security over computing resources.
3. **Application License**[2]
   For developers seeking to develop applications using a licensed model; they would have to declare the purpose or purposes and sector or sectors for which the model would be used. This system would also ensure that modifications to approved AI models remain compliant with safety regulations, particularly when model capabilities are enhanced. Automatic approval would apply to new licensing submissions where they were substantially similar to existing licenses with no significant capability upgrades, though models would still need to seek relevant sector-specific approvals.

[2] Some Congressional leaders have proposed that AI should be regulated via a sector-specific approach. The application license system would enable such an approach by ensuring that unscrupulous AI companies could not evade sector-specific regulation by developing a dangerous general-purpose model, but only seeking regulatory approval for one narrow sector-specific use.

Prohibiting Dangerous AI Capabilities

The regulator would have the power to enforce prohibitions on **specific high-risk AI behaviours**, ensuring that even models operating below regulatory thresholds do not engage in hazardous activities. These would constitute the unacceptable risk thresholds that governments have committed to identify in the Seoul agreement. These prohibited capabilities include:

- No Superintelligent AIs: AI must not surpass human intelligence in general tasks.
- No Unbounded AIs: AI systems should not be developed or deployed *unless* a robust safety case cannot be made regarding their capabilities of concern, ensuring AIs remain predictable and controllable.
- No Environmental Breakout: AI systems must not escape their designated environments or access external systems or networks, even with authorisation, if the regulator deems the degree or scope unsafe by design. (E.G., this ensures that AI models do not pose cybersecurity threats)
- No AIs Improving AIs: AI systems should not improve or develop other AI systems, particularly those not directly written by humans, to prevent runaway AI development out of human control.

Oversight of High-Computation AI models

Through the licensing process, AISI would be able to proactively and iteratively monitor AI model development and deployment, and ensure that the President was fully apprised of all AI capabilities and their implications for US national security.

Establishing Security and Product Safety Standards

To enable lighter-weight regulation, AISI would have the legal authority to establish common standards for AI models' development and deployment while protecting Americans' security and product safety, in line with any other high-potential industry's standards. Such standards would include, but are not limited to, cybersecurity and physical security, insider threat, and reliability standards.

Threat, Scenario, and Trend Assessment

AISI would build in-house capabilities to understand the security and other risks of AI development and deployment, which it could also use to support US IC and DoD analysis.

Governance and Flexibility

The regulator's governance would remain flexible to adapt to future AI developments and risks, with the establishment of an **AI Security Board**, which would be responsible for defining key regulatory thresholds and capabilities for licensing requirements, and have the power to **order the shutdown of dangerous AI models or applications.** Such a board would be staffed by experts and also solicit industry, IC, and DoD perspectives.

A newly created **Scientific Advisory Group** would provide expert input on emerging AI capabilities, risks, and product safety measures. This advisory group, drawing from the best experts at world-leading American institutions like MIT, Harvard, Yale, Stanford, and others,

would work closely with the Board to ensure that regulatory decisions are scientifically informed and aligned with best-practice standards.