A New Regulator for Advanced AI Systems

The Problem – National and Global Security Risks from AI Development

AI that is capable of surpassing human intelligence across many domains is being developed rapidly and without adequate control. Nobel Prize winners, leading AI scientists, and even CEOs of AI companies have warned that advanced AI poses an extinction risk to humanity. Current national legislation lacks binding, enforceable regulations to manage these risks, making it urgent to establish a framework for oversight over powerful AI development.

At the first AI Safety Summit in Bletchley Park in 2023, 28 governments, including the USA, the EU, and the UK, agreed to intervene in AI development posing severe risks, and secured commitments from leading companies to restrict development in cases of unacceptable risks. However, the AI sector remains unregulated and company commitments remain voluntary.

The Solution – Establishing Independent AI regulators

In 2024, the second AI Safety Summit led to an international commitment to build and empower AI Safety Institutes (AISI) to advance our understanding of advanced AI and enable its governance. This followed the lead of the UK and US Governments, who opened national AISIs. We propose **establishing these AISIs as independent national AI regulators**, with the authority to regulate, oversee, and enforce safety standards for frontier AI models. These regulators would ensure that companies developing AI models above certain compute thresholds and general intelligence benchmarks comply with rigorous safety protocols. This would allow countries to harness the benefits of practical AI while mitigating risks posed by the uncontrolled development of superintelligent AI.

Key aspects of the AI regulator mandate would include:

- Licensing frontier AI developers to ensure AI models are safe 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 safety standards for the design, development, deployment, and monitoring of AI systems.

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 licences ensure that only AI developers and operators that meet safety requirements can proceed with their work.

1. **Training Licence**
   For AI developers aiming to train models that exceed a set computational power threshold, set at 10^25 FLOP. Applicants must present detailed risk mitigation plans matching the commitments made at the Seoul AI Summit, including shutdown procedures for AI systems that pose unacceptable levels of risk.
2. **Compute Licence**
   Required for cloud service providers and data centres operating above 10^17 FLOP/s. The compute licence 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 Licence**
   For developers seeking to develop applications using a licensed model. It would ensure that modifications to approved AI models remain compliant with safety regulations, particularly when model capabilities are enhanced. Automatic approval would apply to applications with no significant capability upgrades.

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. If a developer breaks these prohibitions, the regulator is empowered to take away their licenses, delete their AI systems, and even criminally pursue them. These prohibited capabilities include:

- No Superintelligent AIs: AI systems must not surpass human intelligence in general tasks.
- No Unbounded AIs: AI systems for which there are no safety cases (robust arguments for why the AI won't use the capabilities of concern) should not be developed or deployed, ensuring they remain predictable and controllable.
- No Environmental Breakout: AI systems must not be able to escape their designated environments or access external systems or networks, even with authorisation, if the regulator deems the degree or scope unsafe by design.
- No AIs Improving AIs: AI systems should not improve or develop other AI systems, particularly those not directly written by humans.

Governance and Flexibility

The regulator's governance would remain flexible to adapt to future AI developments and risks, with the establishment of an **AI Safety Board**, which would be responsible for defining key regulatory thresholds and capabilities for licensing requirements, and have the power to **order the shutdown of AI models or applications.**

A newly created **Scientific Advisory Group** would provide expert input on emerging AI capabilities, risks, and safety measures. This advisory group would work closely with the Board to ensure that regulatory decisions are scientifically informed and aligned with global safety standards.