
Safety cases and Licensing: A deeper look

[May Dixit](#)

Independent / ERA
Cambridge 2025

With
ControlAI & Apart Research

Introduction

This report evaluates A Narrow Path Phase 0 policies, with the aim of making it more robust and easier to implement. We look in depth at the Safety Cases policy, as well as the Licensing policy to model possible fail points, and suggest some updates that could be useful.

The Results section presents potential challenges in the implementation of the letter and the spirit of these policies. It is broken down into evaluation specific challenges, other safety case specific challenges and licensing related challenges. The report ends with suggestions on additions or updates to the policy to make it more robust.

Methods

In this analysis, I use the **threat modeling methodology and implementation feasibility technique** to ask the following questions [questions brainstormed with Gemini]:

- Identify potential parties that would have incentives to circumvent the spirit of the policy and the loopholes they could use to do it.
 - a. Who has such incentives?
 - b. What are their capabilities?
 - c. What workaround could they use?
- Propose possible updates to the policy that would make these workarounds ineffectual.
- Identify potential hurdles to the implementation of the policy.
- Propose solutions to the technical implementation questions.

Results

Challenges in evaluating and monitoring AI systems.

1. **Difficulty to trust evaluations of non-deterministic models:** General AI models are inherently non-deterministic, as opposed to the deterministic industrial processes where we apply safety cases – nuclear, chemical etc. This poses a unique problem of not being able to trust any benchmarks fully – as the performance may change with the same model in another run or another context.
2. **Context aware models:** It has been shown that models are able to determine if they are being evaluated, which poses a challenge to being able to trust the evaluations.
3. **Defining the right capabilities that will have real world consequences:** Capabilities, themselves, are hard to define in a general intelligence system. There is no clear mapping of specific capabilities to the potential real world consequences. There is no clarity on how they are represented in the internals of the model, and if we are even thinking of them in the right terms.
4. **Scaffolding and Agency:** Safety cases might prove that a model on its own is below threshold. However, with scaffolding, agentic systems and in interaction with a larger environment, the possibilities increase. Think of the case when cyber attackers could get a model to develop the steps to create malware with one unsafe model, and get a safe model to generate each of the separate steps.

Other challenges that are safety case specific:

1. **Scalable safety cases:** Safety cases will need to be scalable to future systems. We might develop a safety case that could be inadequate as model training methods improve. Models may get better at tricking the safety cases instead of actually being below threshold.
2. **False sense of safety:** Safety cases may be used as a false sense of safety while hiding capabilities of the models being tested. Especially, if it is in the interest of the company developing the models to deploy a stronger model while pretending to be compliant.
3. **Talent for auditor roles:** Regulators and auditors will need to be high context and technically minded to understand the nuances in model training and evaluations. There must be sufficient funds allocated to this to attract talented professionals to these roles.
4. **Third party access:** API access will not be enough for auditors to inspect capabilities, especially because even a slight difference in the system prompt or scaffolding would change the behavior of the system significantly. We will need to create privacy preserving, secure methods of auditing models.
5. **Overwhelm leading to low-quality:** We will need to devise a large number of safety cases as there would be numerous capabilities that could possibly lead to disaster. This may create an overwhelm for auditors as well as the

developers resulting in lowered quality. Even if we were to automate the creation of these, the methodology must be heavily scrutinized.

6. **Password locked models:** Models could be trained to be password locked so that while in evaluation they behave with lower capability than in deployment when given the password. This has been proven as a possibility.

Challenges in training / compute license implementation:

1. **Difficulty in monitoring compute unit usage:** Auditing which compute unit was used for which purpose is difficult, especially when companies run parallel experiments and the compute is usually shared between jobs.
2. **They could also train the model in parts** – in a mixture-of-experts fashion, to be put together for a higher capability.
3. **Models arranged in smaller units leading to general intelligence:** Even if this were difficult, narrow models can all be set up in an orchestrator-and-workers framework where the orchestrator may only need to have some planning capabilities, and triggered responses based on the workers' behaviors.

Additional findings

- **Internal deployment:** One thing missing in this policy proposal was internal deployment contexts. In internal deployment, the models used could steer research directions, persuade the employees for some hidden goal or influence the direction of the company in general. Models must be tested before internal deployment and safety cases provided for them as well. However, this brings up the question of how to do this without infringing on the privacy protection of companies.

Suggestions to address some of the challenges discussed

- **Evolving and Adaptive Benchmarks:** Benchmarks must be dynamic and continuously evolve rather than remaining static. They should be designed to be difficult for models to detect as evaluations, which calls for evaluations in more naturalistic and complex settings. This requires ongoing research into novel evaluation methodologies that can adapt to rapid advancements in AI capabilities.
- **Mandatory Aggregated Monitoring Information Sharing:** Companies should be obligated to share **aggregated monitoring information** to ensure transparency and oversight, while still safeguarding proprietary and privacy-sensitive data. This aggregated data can provide regulators with insights into model behavior and performance trends without revealing the specifics of individual data points or trade secrets.
- **Internal Deployment Oversight:** The policy must explicitly include requirements for testing and providing safety cases for **models deployed in internal company contexts**. While respecting corporate privacy, mechanisms must be developed for external auditors to verify internal model safety, potentially through secure, privacy-preserving auditing tools and protocols. This is crucial given the potential for internal AI systems to

subtly influence research directions, employee behavior, or company strategy.

- **Investment in Foundational AI Safety Research:** A significant increase in investment and prioritization is needed for research into **interpretability, formal methods, and general safe AI principles**. While perhaps outside the immediate scope of "A Narrow Path Phase 0" policy, this foundational research is critical for enabling truly safe AI transformation in the long term. This includes developing tools to understand model internals, creating verifiable safety guarantees, and exploring new paradigms for AI design that are inherently safer.
- **Developing Robust Auditing Mechanisms Beyond API Access:** Given that simple API access is insufficient for comprehensive auditing, the policy should mandate the development and adoption of **privacy-preserving, secure methods for auditing models**. This could involve cryptographic techniques, or standardized, secure environments for model inspection that protect sensitive data while allowing regulators deep insight into model behavior and capabilities.
- **Cultivating a High-Context Regulatory Workforce:** To address the need for technically minded regulators and auditors, substantial funding and strategic initiatives are required to **attract, train, and retain top talent** in these specialized roles. This includes establishing dedicated educational programs, competitive compensation, and clear career paths within AI regulatory bodies to ensure that oversight is both informed and effective.
- **Streamlining and Automating Safety Case Development with Scrutiny:** While the sheer volume of necessary safety cases poses a challenge, research and development should focus on **automating their creation and verification**, but with extreme scrutiny of the underlying methodologies. This involves developing frameworks that can generate relevant safety cases based on defined capabilities and potential risks, and tools to automatically check for compliance, reducing manual burden while maintaining high standards of rigor.

Conclusion

This report highlights critical implementation hurdles for AI safety policies, particularly concerning the non-deterministic nature of advanced AI, the challenges in defining its capabilities, and the limitations of current safety cases. Key concerns also emerge around effectively auditing compute usage and the overlooked risks of internal AI deployments. To build genuinely robust AI governance, policies must embrace **adaptive evaluations**, ensure **transparent yet privacy-preserving data sharing**, extend oversight to **internal AI applications**, and significantly invest in **foundational AI safety research and regulatory expertise**. Addressing these points is crucial for policies to keep pace with AI's rapid evolution and effectively manage its associated risks.