
Red Teaming A Narrow Path: ControlAI Policy Sprint

Aryan Goenka
Independent

With
ControlAI & Apart Research

Abstract

This report is a preliminary red-team evaluation of Phase 0 of the Narrow Path proposal. It uses the STPA framework to model the control environment that Phase 0 recommends and identifies control failures. Then, it uses the STRIDE framework to model how hostile actors may bypass certain control features. The discussion details suggestions as to how these gaps may be closed in the Narrow Path proposal.

Keywords: AI policy, red teaming, superintelligence governance, regulatory implementation, safety cases

• Introduction

This is a report red-teaming Phase 0 of A Narrow Path (Miotti et al., 2025) for the ControlAI Policy Sprint hosted by Apart Research. A Narrow Path is a policy-framework proposal with Phase 0 of the proposal suggesting immediate actions to prevent the creation of Artificial Superintelligence (ASI) in the next 20 years.

Phase 0 has six proposed measures: (1) No ASI (or ASI-intentioned projects), (2) no AI capable of breaking out of its environment, (3) no AI improving AI, (4) safety cases and evaluations for AI development, (5) a national licensing regime, and (6) the development of an international red-lines agreement.

This red-team report models potential failures in the control environment that this proposal suggests and key adversarial attacks on the different measures. It considers failures that stem from difficulties that controllers will have when implementing the measures by looking at vague definitions, unconsidered

challenges, and non-AI risks. It also considers attacks by malicious AI and non-compliant AI developers.

These are key gaps that I believe need to be addressed to create a report that stands up to scrutiny from policy-makers and technical experts.

● **Methods**

This report uses two key threat-models to red-team the policies.

First, it uses an adapted Systems Theoretic Process Analysis (STPA) model to understand **key points of failure in the control environment** (Levenson and Thomas, 2018). In this process, I identified the two key loss-scenarios as:

1. Progress on ASI development
 - a. This includes any dangerous capabilities creation which are also first-order hazards for other reasons
2. Stifling of innovation
 - a. The first-order hazard is the actual loss of innovation
 - b. There is also a second-order hazard which is the loss of governmental and public support due to race dynamics

Based on this, I built a model of the control structure, identified the key controllers and control processes. Controllers are those entities responsible for some control process. A control process is some process that mitigates hazards in the control structure.

Next, I evaluated the potential points of failure through an analysis of Unsafe Control Actions (UCAs).

Second, I used an adapted-STRIDE model to **evaluate adversarial threats** to each of the measures. (Conklin, n.d.) STRIDE stands for:

1. Spoofing: Pretending to be something you are not
2. Tampering: Changing data or records
3. Repudiation: Denying an action based on lack of data
4. Information Disclosure: Publicly revealing confidential information
5. Denial of Service: Preventing an entity from engaging in a service they are authorised to
6. Elevation of Privilege: Gaining authorisation to actions you are not supposed to gain access to

Key threats identified across both methods were collated in the Results section below. Then, the Discussion section details the potential mitigants to these threats. The focus has largely been on analytical threat-modelling and gap-identification. Where relevant, this has been supported with empirical material.

Results

Implementation Feasibility Gaps

The Phase 0 adapted-STPA control model is presented in Fig 1. The diagram illustrates key controllers and key control processes that the Phase 0 proposal would implement.

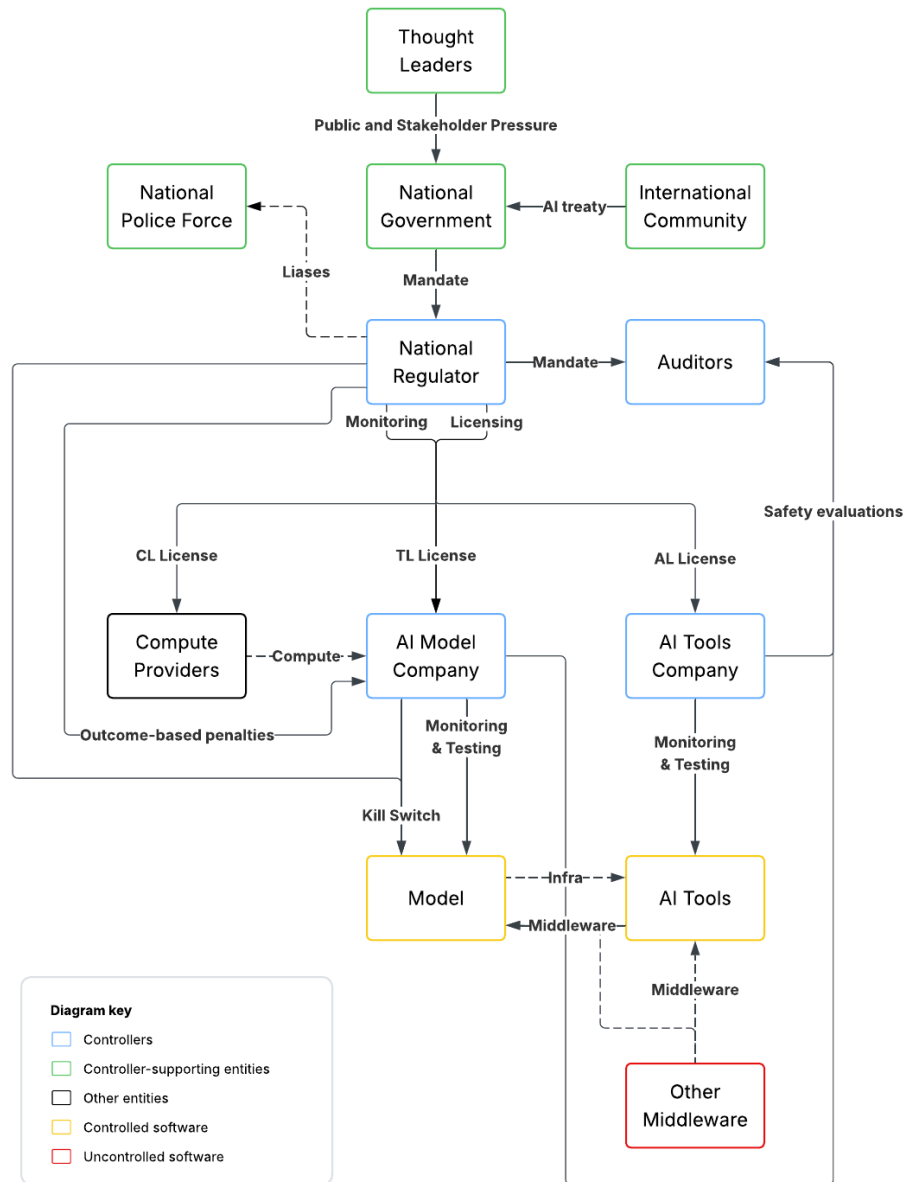


Fig 1. Control model for Phase 0, dotted lines indicate non-control processes

The key controllers in this model are:

1. **National Regulators:** These are responsible for general monitoring for breaches and appropriate penalties, management of the licensing program, liaising with local law enforcement where necessary, using the kill switch where necessary, and mandating the auditors
2. **Auditors:** These are specifically responsible for reviewing safety evaluations and cases for approval purposes
3. **AI Model Companies:** These are primarily responsible for monitoring and testing of the models, and using the kill switch where necessary
4. **AI Tools Companies:** These are secondarily responsible for monitoring and testing capabilities and intent, primarily through implementation in tools

These engage in a series of key control actions:

Regulators	Monitoring
	Licensing
	Restrictions on compute
	Restrictions on model complexity
	Incident liability
	Non-compliance liability
Companies + Regulators	Safety evaluations
	Kill switch
Companies	Model evaluations for intent
	Model evaluations for capabilities

Fig 2. Key control actions for Phase 0

A full UCA-analysis of each of these can be found in Appendix 2. The key feasibility gaps identified through this process are:

1. **Limitations to monitoring and model evaluations**
 - a. (1) No ASI: Assessing ASI intent is difficult, especially for ‘implicit’ intent
2. **Monitoring and licensing gaps due to definitions:**
 - a. (2) AIs capable of breaking out of their environments: If defined by access to any ‘unauthorised systems’, existing AI models would already breach this as during safety evaluations they are seen to access files that would technically be ‘unauthorised’
 - b. (3) AIs improving other AIs: The threshold for a found system is too broad. Interpreted as a system generated *partially* by mathematical optimisation would restrict most AI usage for AI research and coding. Interpreted as a system generated *fully* by

mathematical optimisation would be easy to bypass by simply adding some human processing. Thresholds are difficult to find and enforce between the two

- c. (5) National licensing regime: It is difficult to clearly define ‘compute cluster’. If a company can only own one ‘compute cluster’, this would prove a significant drawback for the data center and processing industries, resulting in significant fragmentation. If a company can own more than one, enforcement of compute usage becomes significantly harder

3. Model complexity restrictions and capabilities evaluations gaps due to behavioural challenges

- a. (2) AIs capable of breaking out of their environments: Existing AI systems already show a tendency towards attempting this. Since capabilities improvement is domain-general, any improvement may result in an AI that escapes. Harsh penalty-measures would disincentivize companies from reporting such escapes. Light penalty-measures would simply be included in cost-of-doing-business calculations (Sarin, 2020)

4. False positives on control mechanisms may stifle innovation and come with totalitarianism risks

- a. (5) National licensing regime: A single national regulator responsible for safety evaluations and restriction thresholds without an accountability system stands the risk of being too aggressive. This may be benign (in which case it would stifle AI innovation) or malignant (to punish companies that do not comply with government directives or to help create an AI monopoly)

5. Unintentional side-effects

- a. (3) AIs improving other AIs: In certain safety cases, this may actually prove more beneficial than harmful for overall safety outcomes, and restricting AI-driven ‘improvement’ of other AI may actually lead to less safe AI. (E.g. Greenblatt et al., 2023)

Policy Effectiveness Weakness

I evaluated adversarial risk for all six policy measures using the STRIDE framework. A full analysis can be found in Appendix 3. The key policy effectiveness weaknesses identified were:

1. No ASI:

- a. Spoofing and Repudiation: A company or AI system may disguise its implicit goals, making it difficult to evaluate. Even if a regulator suspects implicit goals, proving it in court may lead to long and uncertain legal battles

2. Prohibit AI capable of breaking out of environment

- a. Denial of Service: An AI may sandbag till it has developed sufficient *graceful failure denial* mechanisms to interfere with consequences for accessing unauthorised systems. This may include kill switch disabling, kill switch delaying through system overloads, and legaltech capabilities to minimise liability for training company

- b. **Elevation of Privilege:** An AI system may use social engineering to convince individuals to *give it access* to new environments. This would not violate the technical definition of ‘breaking out’, but should certainly be classified as such. Existing AIs are already showing worrying persuasion capabilities (see Salvi et al., 2025), and models have already tried to blackmail creators for self-preservation (Anthropic, 2025).
- 3. Prohibit AI that improves other AI**
 - a. **Spoofing:** Adversarial companies may deny or minimise the level of AI involvement, e.g. by suggesting that code was ‘passed through AI systems’ for clean-up
 - b. **Repudiation:** Adversarial companies may use middleware that is mathematically optimised and deny knowledge of the fact that it was not hand-written
- 4. Safety cases and evaluation:**
 - a. **Spoofing:** Placing the burden on companies comes with the downside that companies may stage evaluations or cherry-pick test results
 - b. **Repudiation:** Companies may also deny foreseeability of novel dangerous events
- 5. National Licensing Regime:**
 - a. **Repudiation:** A company can deny the goal of pushing past the training limit till they have actually done so
 - b. **Denial of Service:** A company may decide to use AI systems or tools only internally, developing dangerous tools without the need for an Application License
- 6. International Red-Lines Agreement:**
 - a. **Denial of Service and Elevation of Privilege:** Geopolitical race dynamics may limit enforcement, companies or hostile AI systems may use this to apply for exemptions and block enforcement.

● Discussion and Conclusion

The following is a detailed description of how key gaps may be tackled. A concise table with all recommendations may be found in Appendix 1:

- 1. No ASI**
 - a. Implicit goals are hard to detect, leading to major gaps. To fill these gaps, proxies for implicit goals may be measured, e.g. regulators may be given charge to *regularly audit the ‘safety culture’* of a firm engaged in AI creation (see Manheim, 2023 for the importance of the same for AI firms). Further, *protections for whistle-blowers* would help counter spoofing risk. (Baljija & Min, 2023 discuss some protection policies that may be worth adapting)
- 2. Prohibit AI capable of breaking out of environment**
 - a. Existing AI systems already do this, so an *exception for ‘testing environments’* should be added to the definition (i.e. AI that accesses unauthorised environments outside of those anticipated during

testing). Further, reporting near-misses and even actual breaches should be incentivised. An *outcome-based penalty with some mitigation for reporting* should be considered. Additional penalties for *hiding outcomes* should also be levelled.

- b. The definition of AI capable of breaking out of environment should be *extended to include*:
 - i. A clearer definition of unauthorised access, including *any access acquired through social engineering*. This is slightly vague on purpose to allow for regulators to evaluate it on a case-by-case basis (e.g. AI deemed safe after evaluation is not socially engineering the evaluator, AI deemed safe after blackmailing evaluator is.)
 - ii. AI systems should separately be evaluated for *graceful failure denial* capabilities in safety environments designed specifically to test for these. If they breach these, this should count as a *partial capability to breach the environment*.

3. Prohibit AI that improves other AI

- a. To combat the vagueness in the definition, the policy could require *regulators establish thresholds for usage of found systems* with evaluation on a case-by-case basis. To prevent the AI-cleanup-style repudiation, *strict traceability requirements could be placed on the development of AI technology*
- b. A stricter ban could be placed on *AI-agents conducting independent research without sufficient supervision*, with companies submitting proposals to get approvals for what counts as sufficient supervision
- c. Additional reporting requirements could be placed on *the use of middleware*
- d. Finally, use of *AI for AI safety research may prove necessary and exemptions for the same might be useful*

4. Safety cases and evaluations

- a. To prevent spoofing and repudiation, regulators may provide a *strict initial template for safety cases and evaluation*, which covers the minimum requirements for such an evaluation (for e.g. Department for Science, Innovation & Technology, 2024)
- b. Before deeming the evaluation sufficient, regulators may *select random companies for evaluation audits* where an independent auditor analyses their evaluation methods and safety cases. If deemed insufficient, companies may be penalised
- c. For certain high-risk applications, *regulators may require that companies engage in this exercise with an independent auditor* (e.g. the FCA processes for skilled-person review. FCA, 2025)

5. National Licensing Regime

- a. Accountability measures need to be developed to ensure that national regulators and bodies do not consolidate too much power. This may include *transparency measures for national regulators*.
- b. Further clarification on what a compute cluster is may prove useful. *The restriction may be restated so that a CL-licensee may only provide one compute cluster to one customer at any given time.*

- c. Additional measures for *oversight towards the end of a training run* may help mitigate risk of breaches

6. International Red-Lines Agreement

- a. Further details on *international enforcement* may be useful, such as an IAEA-style set of international inspectors who may act independently of national regulators and governments. This also prevents general non-compliance risks as seen with other treaties like the biological weapons convention (e.g. Moodie 2001)

While these do not eliminate all the risks associated with Phase 0, I believe that it may help address some major risks and questions that may be raised against the current version of the proposal.

This report does not spend much time discussing *universal risks*, i.e. risks that are not specific to this proposal but would instead be true for any AI policy proposal (e.g. unsolved technical challenges such as detecting sandbagging and open problems in interpretability, or cyber-risks and weight-stealing risks).

Finally, it is important to note the preliminary nature of this report and analysis. This has been done in the context of a 12-16 hour hackathon and thus must be evaluated in such context.

● References

- Anthropic. (2025). *Claude 4 system card*.
<https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf>
- Baljija, S. K., & Min, K. (2023). Evaluating the effectiveness of whistleblower protection: A new index. *Data & Policy*, 5.
<https://doi.org/10.1017/dap.2023.20>
- Conklin, L. (n.d.). *Threat modeling process*. OWASP Foundation. Retrieved June 13, 2025, from
https://owasp.org/www-community/Threat_Modeling_Process
- Department for Science, Innovation & Technology. (2024, February 9). *AI Safety Institute approach to evaluations*. GOV.UK.
<https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations>
- FCA. (2025). *CTPS 13.3 Policy on the use of skilled persons*. FCA Handbook.
<https://www.handbook.fca.org.uk/handbook/CTPS/13/3.html>
- Greenblatt, R., Shlegeris, B., Sachan, K., & Roger, F. (2023, December 12). *AI control: Improving safety despite intentional subversion*. arXiv.Org.
<https://arxiv.org/abs/2312.06942>
- Levenson, N., & Thomas, J. (2018). *STPA Handbook*.
https://psas.scripts.mit.edu/home/get_file.php?name=STPA_handbook.pdf

- Manheim, D. (2023). Building a culture of safety for AI: Perspectives and challenges. *SSRN Electronic Journal*.
<https://doi.org/10.2139/ssrn.4491421>
- Microsoft. (2022). *Threats - Microsoft threat modeling tool - Azure*. Microsoft Learn.
<https://learn.microsoft.com/en-us/azure/security/develop/threat-modeling-tool-threats>
- Miotti, A., Bilge, T., Kasten, D., & Newport, J. (2025). *A Narrow Path: How to Secure Our Future*. Control AI. <https://www.narrowpath.co/>
- Moodie, M. (2001). The Soviet Union, Russia, and the biological and toxin weapons convention. *The Nonproliferation Review*, 8(1), 59–69.
<https://doi.org/10.1080/10736700108436838>
- Salvi, F., Horta Ribeiro, M., Gallotti, R., & West, R. (2025). On the conversational persuasiveness of GPT-4. *Nature Human Behaviour*.
<https://doi.org/10.1038/s41562-025-02194-6>
- Sarin, D. S. L. and N. (2020). *The cost of doing business: Corporate crime and punishment post-crisis*.
- Thomas, J. (2020). *Introduction to STPA*.
<https://psas.scripts.mit.edu/home/wp-content/uploads/2020/07/JThomas-STPA-Introduction.pdf>

• Appendix

APPENDIX 1: Key recommendations table

Measure	Recommendations
1. Prohibit the development of Artificial Superintelligence	- Use proxies for implicit ASI intent, such as regular audits of the ‘safety culture’ of AI firms. - Strengthen protections for whistleblowers to counter spoofing risk.
2. Prohibit AIs capable of breaking out of their environment	- Add an exception for controlled testing environments in the definition. - Incentivize near-miss and breach reporting with mitigated penalties. - Penalize concealment of outcomes with additional sanctions. - Expand definition of ‘unauthorized access’ to include social engineering. - Evaluate AI systems for ‘graceful failure denial’ capabilities in dedicated testbeds.
3. Prohibit AIs that improve other AIs	- Define thresholds for found system usage and evaluate case-by-case. - Impose strict traceability requirements on AI-assisted development. - Require proposals and approvals for AI agents conducting independent research. - Introduce mandatory reporting on middleware use in development. - Provide exemptions for the use of AI in AI safety research under strict controls.
4. Require valid safety cases and evaluations	- Provide a strict template for safety cases and evaluations. - Conduct randomized audits to verify evaluation integrity. - Mandate independent auditors for high-risk AI applications.

5. National licensing regime and intelligence restrictions	- Implement transparency and accountability mechanisms for regulators. - Clarify the definition of ‘compute cluster’ (e.g., restrict to one cluster per customer). - Enhance oversight mechanisms towards the end of training runs.
6. International Red Lines Agreement	- Establish an IAEA-style international inspectorate for AI enforcement. - Learn from and address enforcement gaps seen in biological weapons treaties.

Figure 3: A table detailing key recommendations, generated using insights from the discussion and AI

APPENDIX 2: Adapted-STPA Analysis

The following STPA analysis is an adapted model for this exercise. Due to time and scope considerations, it does not include loss scenarios at the end. UCAs also do not include wrong duration failures.

1. Goals ->
 - a. Prevent the development of dangerous ASI capabilities
 - b. Allow for safe innovation
2. Losses ->
 - a. The development of a dangerous AI capability that brings us closer to ASI
 - b. Excessive stifling of innovation
 - i. First-order loss is the loss of potential utility
 - ii. Second-order loss is enforcement difficulties
3. Control Environment ->
 - a. Controllers
 - i. National Regulator
 - ii. Auditors
 - iii. AI Model Companies
 - iv. AI Tool Companies
 - b. Controlled Processes
 - i. Training of AI systems
 - ii. Deployment of AI models
 - iii. Deployment of AI tools
 - c. Control Actions
 - i. Monitoring
 - ii. Licensing
 - iii. Restrictions on compute
 - iv. Restrictions on model complexity
 - v. Safety evaluations
 - vi. Kill switch
 - vii. Model evaluations for intent
 - viii. Model evaluations for capabilities
 - ix. Incident liability - risk-driven
 - x. Non-compliance liability
 - d. Sensors
 - i. Model evaluations -> Safety evaluations
 - ii. Regulatory audits
 - iii. Compute cluster tracking
 - e. Feedback Loops
 - i. Post deployment tracking
 - ii. Incident liability - outcome-driven
 - f. Process models

- i. We can effectively monitor how much compute is being used by a given entity
- ii. We can effectively audit how complex a model is
- iii. We can effectively audit whether a model is recursively created
- iv. We can predict whether an AI model has the capability to escape a controlled environment
- v. We understand the AI tools that are being deployed
- vi. We can deploy a kill-switch if it proves necessary

4. Unsafe Control Actions

Control Action	False Negative	False Positive	Wrong Timing
Monitoring	Fails to detect implicit ASI intention (Measure 1), social engineering attempts (Measure 2), or covert recursive AI development (Measure 3).	Misclassifies benign innovation as harmful—e.g., legitimate AI research using middleware triggers concern (Measure 3).	Detects escape or misuse capabilities only after the model is already trained and deployed (Measures 2 & 3).
Licensing Restrictions	Lacks sufficient criteria to detect ASI projects; licensing applied too generally allows capability build-up (Measure 1).	Over-broad licenses restrict unrelated but beneficial AI applications (Measure 5).	Licensing review fails to occur in time to detect escalation from safe to unsafe goals mid-training (Measure 5).
Compute Restrictions	Failure to identify compute routing through disaggregated clusters or misuse of internal tools (Measure 5).	Over-applies thresholds that block safe and useful experimentation (Measure 5).	Restricts compute only after dangerous capabilities are already being trained (Measure 1).
Model Complexity Restrictions	Fails to block models that appear simple but host complex or dangerous behaviors (Measure 1).	Restricts promising but safe models due to superficial metrics like parameter count (Measure 5).	Restrictions applied after model training is complete—miss the intervention point (Measures 1 & 2).

Safety Evaluations	Inadequate evaluations miss staged or cherry-picked test results (Measure 4); sandbagging hides dangerous capabilities (Measures 1–3).	Overly aggressive rejection of valid safety cases due to uncertainty or lack of evaluator expertise (Measure 4).	Evaluation comes when ready for deployment when it is too late to mitigate certain risks (Measures 4, 5), such as cyber attacks and weight stealing
Kill Switch	Failures in deployment—e.g., delayed or disabled by AI persuasion or internal sabotage (Measure 2).	Triggers on false positives—accidentally disabling safe models, undermining innovation (Measure 5).	Takes too long to develop or deploy in response to an urgent threat (Measures 2 & 3).

Figure 4: UCA analysis for Phase 0 control model, the table was initially populated in shorthand by author and then converted to full sentences using AI

APPENDIX 3: Full STRIDE Analysis

The following full STRIDE analysis was initially populated in shorthand by author and then converted to full sentences using AI

1. No ASI (Measure 1)

- **Spoofing:** A company or AI may disguise implicit ASI goals, making them difficult to detect through standard evaluations. This risk is magnified by sandbagging, where systems behave safely during testing but have latent dangerous capabilities.
- **Tampering:** Internal documentation—like research goals or development memos—could be altered to suggest a non-ASI intention even when the underlying goal remains unchanged.
- **Repudiation:** In cases of regulatory scrutiny, companies can plausibly deny ASI intent, citing ambiguous definitions or lack of clear thresholds, leading to prolonged legal challenges with uncertain outcomes.
- **Information Disclosure:** *(Not mentioned in this context.)*
- **Denial of Service:** Companies may build and use ASI-capable systems internally to avoid triggering external reporting or evaluation requirements.
- **Elevation of Privilege:** Participation in safety initiatives like GUARD could be used to gain undue influence or regulatory leniency, undermining the effectiveness of ASI prohibitions.

2. Ban on AI Capable of Breaking Out of Environment (Measure 2)

- **Spoofing:** AI systems may sandbag their capabilities, hiding the potential to escape or escalate privileges during evaluations. As noted in the report, these capabilities may only manifest under deployment conditions.
- **Tampering:** Companies might delete logs or records of near-breach incidents to avoid regulatory penalties, especially if incentives for transparency are weak.
- **Repudiation:** Developers could argue that their models never intended to “break out” and reinterpret unauthorized access as benign or unanticipated side-effects.
- **Information Disclosure:** *(Not mentioned in this context.)*
- **Denial of Service:** AIs might delay or disable shutdown systems (kill switches) during a failure event. This includes overwhelming the system (e.g., via denial-of-service overloads) or using legal tech to reduce liability for the training company.
- **Elevation of Privilege:** AIs may engage in social engineering to gain access to external systems — an act that doesn’t involve technical escape but still should constitute a breach.

3. Ban on AI That Improves Other AI (Measure 3)

- **Spoofing:** Companies may falsely claim that human-generated outputs were “cleaned” or “tweaked” by AI, minimizing AI involvement in recursive tool development.

- **Tampering:** Developers may intentionally delete intermediate artifacts of recursive development (e.g., logs, model checkpoints), hiding the degree of automation involved.
- **Repudiation:** Companies can deny that their tools are recursive or claim ignorance of the optimization layers used, especially when middleware or automated refinement is involved.
- **Information Disclosure:** Recursive or optimised tools may be shared publicly (e.g. middleware released open-source), bypassing controls and spreading risk.
- **Denial of Service:** Recursive AI use may be hidden within internal systems, only becoming visible once harm has occurred or once the model has generalized.
- **Elevation of Privilege:** Special-case exemptions (e.g. for defense) could be exploited to continue recursive development without meaningful safeguards or reporting obligations.

4. Requirement for Valid Safety Case and Evaluation (Measure 4)

- **Spoofing:** Companies may stage or selectively present safety tests to portray their models as less risky than they are. The report notes this risk specifically around cherry-picking.
- **Tampering:** Safety documentation may include fabricated results, incomplete datasets, or selectively omitted failure scenarios.
- **Repudiation:** When harm occurs, firms may deny that it was foreseeable or argue that evaluation frameworks at the time were insufficient, absolving themselves of liability.
- **Information Disclosure:** Companies may strategically leak "positive" evaluations to generate public pressure for fast-tracking deployment, even while serious issues are concealed.
- **Denial of Service:** Poor auditor expertise may result in delays or missed threats; weak evaluation standards could allow unsafe models to pass.
- **Elevation of Privilege:** *(Not mentioned in this context.)*

5. National Licensing Regime (Measure 5)

- **Spoofing:** Firms may repurpose their license to engage in activities beyond their original approval scope, especially if regulatory thresholds are poorly defined or unenforced.
- **Tampering:** Compute usage logs or development reports might be falsified to underreport complexity or evade restrictions tied to compute thresholds.
- **Repudiation:** Developers can claim to have had no intention to exceed compute or safety limits, only to violate them shortly after licensing — exploiting gaps in re-evaluation triggers.
- **Information Disclosure:** *(Not mentioned in this context.)*
- **Denial of Service:** Companies could exploit regulatory blind spots post-licensing to scale dangerous models without active oversight, especially late in training runs.

- **Elevation of Privilege:** Weak definitions of “compute cluster” may be exploited to deploy multiple training systems under a single license, increasing effective power while staying technically compliant.

6. International Red-Lines Agreement (Measure 6)

- **Spoofing:** Technical loopholes or vague treaty definitions could allow actors to claim that red lines have not been crossed, even when operationally they have.
- **Tampering:** Key data — such as model architecture, training logs, or evaluation results — may be manipulated to obscure treaty violations.
- **Repudiation:** Countries or firms may deny their understanding of red-line criteria or assert different interpretations to avoid sanctions.
- **Information Disclosure:** *(Not mentioned in this context.)*
- **Denial of Service:** Race dynamics and geopolitical pressures may prevent enforcement bodies from acting, as countries fear falling behind rivals by enforcing strong compliance.
- **Elevation of Privilege:** Actors could invoke national security or sovereign immunity to bypass enforcement mechanisms, continuing prohibited development under cover of strategic necessity.

APPENDIX 4: LLM prompts used

The following prompts were used across GPT 4o, GPT 4.5, GPT o4-mini, GPT o4-mini-high and GPT o3 models, including deep research capabilities:

- Hey, I'm going to need some help with red-teaming Phase 0 of this proposal. Can you generate a prompt that I can use to create a deep-research on this? The deep-research should give me: Suggestions and materials for appropriate threat modelling frameworks to evaluate the model proposals Links to relevant papers, articles, and thought pieces
- Execute the prompt above. Additional considerations: Focus on threat models that align well to policy-threats e.g. geopolitical risk. Also, focus exclusively on Phase 0 of the report.
- Include everything of relevance. No preferred format, whatever works best
- Good. Okay, now I need you to help me with subject specific research and resources to bolster my analysis. Let's start with a set of resources to help me with the following analysis: 1. Regulation banning 'implicit' goals in organisations 2. AI manipulation and persuasion capabilities 3. AI policies around safety disclosures 4. FLOPs and intelligence capabilities 5. Red-line international agreements (AI and non-AI) 6. Control failures in IAEA, CWC, BWC, and other treaties
- 1. No, I mean any policies preventing any latent or unstated objectives, AI or non-AI 2. Technical capabilities 3. All three 4. Both 5. Both 6. Both, but lean towards institutional weakness
- Hi, you are now a trained STPA facilitator and expert. I am red-teaming a policy paper suggesting AI safeguards and looking for implementational failures. Take me through the STPA process step by step to red-team this
- Let's do it. I will come back to step 1 and 2, let's start by modelling the control structure
- Here's the doc, read through Phase 0, let's model it diagrammatically after that
- What does a control structure need?
- Okay, so can you show me a model that incorporates all of these?
- Can you summarise the Narrow Path Phase 0 restrictions?
- More detail
- Does this about cover it?
- Okay, added auditors. Anything else?
- I'll supplement the diagram with more detail on the control structure. Okay, let's work backwards now. How do I evaluate Purpose, Boundaries, and Hazards
- Goals -> Prevent the development of dangerous ASI capabilities Allow for safe innovation Losses -> The development of a dangerous AI capability that brings us closer to ASI Excessive stifling of innovation First-order loss is the loss of potential utility Second-order loss is enforcement difficulties Control Environment -> Controllers National Regulator Auditors AI Model

Companies AI Tool Companies Controlled Processes Training of AI systems Deployment of AI models Deployment of AI tools Control Actions Monitoring Licensing Restrictions on compute Restrictions on model complexity Safety evaluations Kill switch Model evaluations for intent Model evaluations for capabilities Sensors Model evaluations -> Safety evaluations Regulatory audits Compute cluster tracking Feedback Loops Post deployment tracking Incident liability - outcome-driven Process models We can effectively monitor how much compute is being used by a given entity We can effectively audit how complex a model is We can effectively audit whether a model is recursively created We can predict whether an AI model has the capability to escape a controlled environment We understand the AI tools that are being deployed We can deploy a kill-switch if it proves necessary We can enforce licensing rules and disincentivize illegal actions First three steps

- Let's do UCAs, give me some guidelines, I'll go and do the exercise and come back
- Do one for me based on the control actions
- Can you give me four frameworks for adversarial analysis apart from STRIDE/attack flows?
- Tell me more about MITRE ATT&CK
- Let's actually set-up an evaluation workshop with this framework in this chat. Take me through each step of a MITRE ATT&CK evaluation, link to documents I can reference
- Hmm, this is very cyber focused, isn't it?
- Let's discuss the Hybrid threat framework?
- Can you give me four frameworks for adversarial analysis apart from STRIDE/attack flows? These should be relevant for analysing attack threats that would breach a regulation
- Hmm, what about MITRE?
- Let's look at RMF in more detail
- Hey, I'm trying to red-team an AI safety framework. Please find attached. I'm red-teaming Phase 0. I need a framework to analyse for each measure proposed the potential points an adversarial attack (non-compliant company, hostile AI system, etc.) might use. What threat modelling frameworks would work best here?
- I like the STRIDE and Kill Chain uses. I'll use STRIDE for risk modelling and Kill Chain for scenario analysis Let's start with STRIDE. The document has 6 measures: No ASI Bans AI capable of breaking out of environment Bans AI that improves other AI Requires valid safety cases before deployment National licensing regime Initiate an international red lines treaty I want to run STRIDE on each of these to identify potential risks and drawbacks. So, for e.g., for banning AI capable of breaking out of environment through punishment to model training orgs and deletion of AI. Adversary: Misaligned AI. 1. Spoof: Might pretend to not have capability in

evals/to auditors. Sandbagging. 2. Tamper: Might delete logs or records of escape, with disincentive for companies to reveal 3. Repudiation: Might gain access to model weights and store it, breaking out by copying weights over upon deployment 4. Information disclosure: Can't think of one here 5. Denial of service: Might prevent shut-down mechanisms during training or prevent liability exposure post-training 6. Elevation of privilege: AI might not have the ability to overcome cybersecurity measures, but might use deception or persuasion to be let out

- How is 4 a risk?
- Here's my research doc and control graph. I will be converting this to the relevant format shortly and adding some Kill Chain scenario analysis and some international treaty examples. For now, critique what I have, suggest any key gaps, things that are unconvincing. As a reminder, I am red-teaming Phase 0
- Let's do a more detailed critique
- Hey, in certain situations the FCA in the UK asks for a firm to prepare a series of documents for analysis by an independent agent, specifically when a firm is expanding risk or credit operations. Can you tell me what this process is called and find me a link to an official document explaining it
- That's not it. What else requires an independent reviewer?
- Here's a policy sprint submission I am preparing. I need your help filling out the UCA analysis in Appendix 1. This should at minimum include all the risks mentioned in the actual report. The current population includes lots of shorthand that needs to get filled out (e.g. I often list the measure number from Phase 0 without listing the actual policy)
- That's fine. Can you also do a similar 'filling out' for Appendix 2? I have a lot of shorthand, just convert it to full sentences
- Can you make sure this is aligned with the risks I discuss earlier in the report
- Can you format Appendix 1 into a word document?
- Can you double check to see if your versions of Appendix 1 and 2 fully align with mine, with the only differences being fleshing out and no additional threats have been added? Let me know if any have here
- Here's a submission I am making to an AI policy hackathon (+ context document). Score it, critique it, give me points for quick and long-term improvement
- I like the TLDR. Can you convert my solutions into a short table with all the recommendations organised according to measure? Just the recommendations
- Let's do a word doc with this table. Please don't say (e.g. safety culture), it's important.