

# Red Teaming a Narrow Path: ControlAI Policy Sprint

**Author Name:** Marshall White

**Co-authoring support:** ChatGPT-4o, Perplexity AI (documented in Appendix per LLM usage policy)

**Affiliation:** Student (Independent Contributor), Curtin University (Australia) & University of International Business and Economics (Beijing)

**With:** ControlAI & Apart Research

---

## Abstract

This red team scenario critiques the "No Unbounded AIs" and "No AIs Improving AIs" assumptions embedded in *A Narrow Path* Phase 0 policies. We argue that focusing on controlling training runs and centralising regulatory enforcement fails to account for post-deployment systemic intelligence. Through the hypothetical lens of a multinational corporation—Parallax Industries—we show how AGI-level capabilities can emerge across CBDC-based infrastructure and DAO-linked smart contracts without violating FLOP thresholds or requiring centralised compute. Our findings reveal that the current regulatory proposals are ontologically brittle and functionally exploitable. We recommend redefining ASI based on functional impact and systemic emergence rather than architectural constraints, expanding oversight to economic systems, and treating persistent self-modifying platforms as dynamic ASI-infrastructure. The implications challenge the narrow definition of threat, urging a more systemic and plural approach to safety.

**Keywords:** AI governance, red teaming, superintelligence risk, regulatory frameworks, decentralized AI, central bank digital currency (CBDC), decentralized autonomous organizations (DAO), systemic emergence, post-deployment evolution, compliance evasion

---

## Contents

|                                                                                  |    |
|----------------------------------------------------------------------------------|----|
| Red Teaming a Narrow Path: ControlAI Policy Sprint.....                          | 1  |
| Abstract.....                                                                    | 1  |
| Introduction .....                                                               | 3  |
| Methods .....                                                                    | 4  |
| Expanded Methodological Framework for Red Teaming AI Policy Implementation ..... | 4  |
| Historical Precedent Analysis .....                                              | 4  |
| Comparative Policy Review .....                                                  | 6  |
| Threat Modelling: Multinational Corporations with Sovereign Access .....         | 8  |
| Feasibility Testing: Policy Intent vs. Enforcement Reality.....                  | 9  |
| Circumvention Pathway Identification .....                                       | 9  |
| Scenario Modelling: Agentic Systems and Economic Infrastructure .....            | 10 |
| Network Analysis of Financial Infrastructure Coordination .....                  | 11 |
| Game-Theoretic Modelling of Incentive Structures .....                           | 11 |
| Structured Expert Elicitation for Uncertainty Quantification .....               | 12 |
| Results .....                                                                    | 13 |
| Policy Redteam Results .....                                                     | 13 |
| Compliance Failure Table.....                                                    | 16 |
| Discussion and Conclusion.....                                                   | 16 |
| References .....                                                                 | 18 |
| Appendix.....                                                                    | 21 |
| Prompt logs (see "LLM Prompts Used").....                                        | 21 |
| Perplexity.....                                                                  | 21 |
| ChatGPT 4o.....                                                                  | 25 |

## Introduction

This report investigates whether *A Narrow Path* Phase 0 policies can meaningfully prevent AGI or ASI development over a 20-year horizon, particularly under non-traditional emergence models. We red team the following policies:

- No AIs improving AIs
- No breakout from constrained environments
- No unbounded AIs
- Limit AI generality below superhuman level

Our threat model centers on a well-resourced corporate actor embedded in international fintech infrastructure. The actor does not develop ASI via centralised training, but via strategic deployment of modular, bounded AIs into financial and logistical systems with smart contract-based evolution. The result is a distributed swarm-intelligence that gradually exceeds human-level control in macroeconomic coordination.

This red team contributes to policymaking by illustrating how apparent compliance can mask systemic violation, and how existing enforcement regimes can be coopted to incubate emergent superintelligence.

---

## Methods

The methodological approach taken in this report is tightly aligned with the red teaming objectives of the A Narrow Path policy sprint. Historical precedent analysis drawing on case studies such as the e-CNY, Eurodollar system, Diem project, and shadow banking networks demonstrates how seemingly bounded infrastructures have historically enabled systemic regulatory evasion. Comparative policy reviews draw critical parallels between AI governance and domains like nuclear non-proliferation and dual-use technology control, highlighting jurisdictional blind spots and the limits of enforceability. Threat modelling of a multinational actor with economic leverage directly mirrors real-world risks posed by private sector alignment drift. Feasibility testing addresses the core question of whether proposed enforcement mechanisms can constrain AI-adjacent architectures across global jurisdictions. Finally, scenario modelling explores the novel vector of autopoietic infrastructure systems whose intelligence emerges through post-deployment interaction expanding the frame of AGI risk beyond conventional FLOP-based or model-centric paradigms. Collectively, these methods directly interrogate the effectiveness and resilience of Phase 0 policies under adversarial pressure.

## Expanded Methodological Framework for Red Teaming AI Policy

### Implementation

This expansion provides comprehensive methodological detail for each analytical approach used in red teaming the "A Narrow Path" Phase 0 policies, incorporating established frameworks from regulatory analysis, financial systems research, and threat assessment methodologies.

### Historical Precedent Analysis

#### *e-CNY*

China's e-CNY system illustrates how a centrally designed, compliance-bound infrastructure can still enable persistent optimisation, dynamic feedback integration, and emergent behavioural adaptation. While not currently a threat, its architecture shares critical features with our Parallax scenario: economic embeddedness, incentive-driven usage, and post-deployment evolution without retraining. The e-CNY demonstrates how future AI-adjacent systems could exhibit ASI-like properties without violating FLOP or architectural constraints. Its architecture—highly embedded, policy-reactive, and capable of integration with AI

modules—demonstrates the plausibility of post-deployment systemic intelligence. It serves as a proof-of-concept for how future economic systems could exhibit ASI-like properties without ever violating training run thresholds.

### *Eurodollar Market Development and Regulatory Evasion*

The Eurodollar market serves as a paradigmatic example of how financial innovation can circumvent regulatory oversight through jurisdictional arbitrage. Emerging in the late 1950s as a mechanism for Soviet banks to hold U.S. dollars outside American regulatory jurisdiction, the market demonstrated how actors can exploit regulatory gaps to create parallel financial systems. By the 1960s, London had become the dominant centre for Eurodollar trading, offering favourable time zones, language advantages, and crucially, an absence of Federal Reserve oversight.

Methodological Application: Historical analysis reveals three key patterns relevant to AI policy circumvention:

- (1) regulatory pre-emption through offshore operations,
- (2) rapid scaling once regulatory gaps are identified, and
- (3) institutional momentum that makes retroactive regulation politically and economically difficult.

The Eurodollar market's growth from a niche Soviet banking solution to a \$10 trillion shadow banking system illustrates how initially narrow regulatory loopholes can evolve into systemic vulnerabilities.

### *Facebook's Diem Project: Political Interference vs. Technical Compliance*

The collapse of Facebook's Diem stablecoin project provides critical insights into how political pressure can override technical regulatory compliance. Despite project leadership claims that all regulatory concerns had been addressed by 2021, political intervention from the U.S. Treasury and Federal Reserve ultimately terminated the initiative. Policymakers, particularly in Europe, viewed Diem as a fundamental challenge to monetary sovereignty given Facebook's global reach of 2.9 billion users.

Methodological Framework: The Diem case illustrates the distinction between formal regulatory compliance and political acceptability in AI governance. Analysis focuses on:

- (1) technical compliance achievement versus political viability,

- (2) the role of systemic importance in regulatory response, and
- (3) the use of financial infrastructure pressure (debanking) as an enforcement mechanism.

### *Shadow Banking Systems and Regulatory Arbitrage*

Shadow banking systems demonstrate sophisticated methods of regulatory arbitrage through structural complexity and jurisdictional shopping. Chinese banks, for example, used on-balance sheet shadow loans to window-dress regulatory capital ratios while maintaining actual risk exposure. These institutions successfully manipulated leverage ratios through shadow lending while keeping risk-weighted capital ratios compliant with regulatory requirements.

Analytical Components: Shadow banking analysis incorporates:

- (1) identification of regulatory gaps between functionally similar activities,
- (2) assessment of window-dressing techniques that maintain technical compliance while violating regulatory intent, and
- (3) evaluation of market pricing of hidden risks versus regulatory recognition 57.

## Comparative Policy Review

### *Nuclear Non-Proliferation Verification Systems*

The International Atomic Energy Agency (IAEA) safeguards system provides the most sophisticated existing framework for monitoring dual-use technology compliance 89. IAEA verification relies on multiple complementary approaches: facility design review, nuclear material accounting, regular reporting requirements, and comprehensive inspection regimes including routine, special, and environmental sampling procedures.

Methodological Translation: Nuclear safeguards methodology adapts to AI policy through:

- (1) Facility-Based Oversight - equivalent to compute infrastructure monitoring,
- (2) Material Accounting - tracking AI model development and deployment,
- (3) Environmental Sampling - detecting undeclared AI development through indirect indicators, and
- (4) Containment and Surveillance - continuous monitoring systems with remote data transmission capabilities.

The IAEA system's evolution from facility-specific monitoring (INFCIRC/26) to comprehensive safeguards covering entire national nuclear programs (INFCIRC/153) provides a framework for scaling AI oversight from individual systems to national AI ecosystems.

### *Dual-Use Export Controls and Implementation Challenges*

Dual-use export control regimes face significant implementation challenges that parallel AI policy enforcement difficulties. The European Union's export control framework identifies key shortcomings: decision-making blockages within multilateral regimes, increasing use of unilateral controls outside established frameworks, lack of unified EU approach, and insufficient coordination of national control lists.

Policy Framework Analysis: Dual-use controls reveal four critical enforcement vulnerabilities applicable to AI governance:

- (1) Multilateral Coordination Failures - individual members can block comprehensive controls,
- (2) Unilateral Bypass Mechanisms - actors can circumvent multilateral agreements through bilateral arrangements,
- (3) Jurisdictional Inconsistency - regulatory arbitrage opportunities across different national frameworks, and
- (4) Technology Evolution Speed - regulatory frameworks lag behind technological development 10.

### *Financial Regulatory Evasion Mechanisms*

Financial regulatory arbitrage provides direct methodological parallels for AI policy circumvention analysis. Regulatory arbitrage exploits gaps between economic substance and legal treatment, utilising three primary mechanisms: jurisdictional arbitrage (exploiting regulatory differences between countries), institutional arbitrage (restructuring legal entities to avoid specific regulations), and product arbitrage (financial engineering to achieve identical outcomes through different regulatory pathways).

Methodological Framework: Financial evasion analysis incorporates:

- (1) Differential Cost Analysis - quantifying regulatory compliance costs across jurisdictions,

(2) Functional Equivalence Testing - identifying activities that achieve identical outcomes through different regulatory treatments, and

(3) Innovation Response Assessment - evaluating how regulatory constraints drive financial product innovation designed to maintain regulatory compliance while achieving prohibited objectives.

## Threat Modelling: Multinational Corporations with Sovereign Access

### Sovereign-Level Capabilities Assessment

Modern multinational corporations, particularly in financial technology, possess capabilities that approach sovereign-level influence over critical infrastructure. Digital sovereignty frameworks recognise that certain private actors can exercise control over data processing, operational infrastructure, and software systems at scales comparable to national governments.

Threat Assessment Framework: Corporate sovereign access evaluation examines:

- (1) Infrastructure Control - ability to influence or control critical economic systems,
  - (2) Jurisdictional Leverage - capacity to exploit regulatory differences across multiple jurisdictions simultaneously,
  - (3) Regulatory Capture Potential - resources and incentives to influence policy development, and
  - (4) Systemic Importance - economic entrenchment that creates "too big to regulate" dynamics
- 1415.

### *Capability-Motivation Matrix Development*

Corporate threat modelling employs structured assessment of both technical capabilities and economic motivations. Agentic AI systems deployed through corporate infrastructure can achieve coordination capacities that exceed traditional centralised AI development while maintaining regulatory compliance at the component level.

Assessment Methodology: Threat modelling incorporates:

- (1) Technical Capability Mapping - assessment of distributed AI deployment capacity through existing infrastructure,

(2) Economic Incentive Analysis - evaluation of financial motivations for regulatory circumvention,

(3) Resource Availability Assessment - corporate financial and technical resources available for sustained regulatory arbitrage efforts, and

(4) Risk Tolerance Evaluation - corporate willingness to accept regulatory and reputational risks.

## Feasibility Testing: Policy Intent vs. Enforcement Reality

### Enforcement Mechanism Analysis

AI policy enforcement faces unique challenges due to the distributed nature of AI development and deployment. Comparative analysis of enforcement mechanisms reveals significant gaps between policy intention and practical implementation capabilities.

Enforcement Framework Assessment: Feasibility testing evaluates four primary enforcement approaches:

(1) Real-Time Automated Conformity Assessment - technical systems for continuous compliance monitoring,

(2) Standardisation and Certification - formal compliance verification procedures,

(3) Algorithmic Impact Assessment - evaluation of AI system effects post-deployment, and

(4) Algorithmic Auditing - comprehensive examination of AI system compliance 17.

Multi-criteria decision analysis (MCDA) of these enforcement mechanisms indicates that external compliance mechanisms (implemented by independent bodies) are more effective but significantly more difficult to implement than internal corporate compliance systems 17.

## Circumvention Pathway Identification

Regulatory circumvention analysis draws from established patterns in financial regulatory arbitrage and dual-use technology control evasion. Circumvention pathways typically exploit three fundamental gaps: definitional ambiguities in regulatory frameworks, jurisdictional inconsistencies in enforcement, and technological evolution that outpaces regulatory adaptation.

Circumvention Analysis Framework: Systematic identification of potential bypass mechanisms includes:

- (1) Legal Gap Exploitation - identification of definitional loopholes and regulatory uncertainties,
- (2) Jurisdictional Arbitrage Assessment - mapping regulatory differences across relevant jurisdictions,
- (3) Technical Workaround Development - evaluation of methods to achieve prohibited outcomes through compliant means, and
- (4) Enforcement Capacity Limitations - assessment of practical constraints on regulatory oversight.

## Scenario Modelling: Agentic Systems and Economic Infrastructure

### Monte Carlo Simulation for Regulatory Scenarios

Monte Carlo simulation provides a robust framework for modelling the probability distributions of different AI governance outcomes under uncertainty [19,20]. This methodology is particularly applicable to AI policy analysis due to the inherent uncertainty in technological development trajectories and regulatory responses.

*Simulation Framework: Monte Carlo modelling for AI governance incorporates:*

- (1) Variable Range Definition - establishing probability distributions for key uncertain parameters such as technological development speed, regulatory response times, and enforcement effectiveness,
- (2) Scenario Generation - running thousands of simulations with different combinations of variable values,
- (3) Outcome Analysis - statistical evaluation of policy effectiveness across different scenarios, and
- (4) Risk Quantification - assessment of probabilities for various policy failure modes.

Applications include quantifying the probability of achieving specific AI safety objectives, measuring the likelihood of regulatory capture, and assessing the risk that distributed AI systems will exceed human control thresholds.

## Network Analysis of Financial Infrastructure Coordination

Network analysis provides essential tools for understanding how distributed AI systems might achieve coordination through existing financial infrastructure. Financial networks exhibit complex adaptive system properties where emergent behaviours arise from the interactions of individual components.

Network Modelling Framework: Infrastructure coordination analysis employs:

- (1) Centrality Analysis - identification of critical nodes and links in financial networks that could serve as coordination points for distributed AI systems,
- (2) Systemic Risk Assessment - evaluation of cascade failure potential when AI systems are embedded in interconnected infrastructure,
- (3) Emergence Pattern Detection - identification of how local AI behaviours might aggregate into system-wide intelligence, and
- (4) Critical Path Analysis - mapping of essential communication and coordination pathways.

Financial markets demonstrate emergent properties including volatility patterns, risk distributions, and coordination behaviours that exceed the sum of individual participant actions <sup>24</sup>. This framework directly applies to assessing how distributed AI systems might achieve coordination capabilities that exceed regulatory thresholds.

## Game-Theoretic Modelling of Incentive Structures

Game theory provides essential frameworks for modelling strategic interactions between AI developers, regulators, and other stakeholders under competitive conditions. Incentive structures embedded in regulatory frameworks create strategic games where actors optimise their behaviour based on expected payoffs and the anticipated actions of other players.

Game-Theoretic Analysis Components: Strategic modelling incorporates:

- (1) Payoff Matrix Development - quantification of costs and benefits for different strategic choices by various actors,
- (2) Equilibrium Analysis - identification of stable strategic configurations where no actor benefits from unilateral strategy changes,
- (3) Evolutionary Dynamics - modelling of how strategies evolve over time as actors learn and adapt, and

(4) Mechanism Design - evaluation of how regulatory structures can be designed to align individual incentives with collective objectives.

Applications include modelling the dynamics between AI developers seeking to maximise capabilities while minimising regulatory costs, regulators attempting to balance innovation promotion with safety assurance, and international coordination challenges where countries face incentives to defect from cooperative AI governance agreements.

### Structured Expert Elicitation for Uncertainty Quantification

Structured Expert Elicitation (SEE) provides systematic methods for incorporating expert judgment into decision-making under uncertainty, particularly valuable when empirical data is limited or unavailable. This methodology is especially relevant for AI governance where technological trajectories and regulatory effectiveness remain highly uncertain.

Expert Elicitation Framework: SEE implementation includes:

- (1) Expert Selection and Calibration - identification and validation of subject matter experts across relevant domains including AI development, regulatory implementation, and technology policy,
- (2) Question Design and Elicitation Protocols - development of structured procedures for extracting quantitative assessments from expert knowledge,
- (3) Aggregation Methods - systematic combination of individual expert judgments into collective assessments, and
- (4) Uncertainty Characterisation - quantification of confidence intervals and disagreement levels among experts.

For AI policy analysis, SEE enables quantification of key uncertainties including timelines for AI capability development, effectiveness of different regulatory approaches, likelihood of successful international coordination, and probability of various failure modes in AI governance systems.

This comprehensive methodological framework provides the analytical foundation for rigorous red team assessment of AI policy proposals, incorporating established techniques from multiple disciplines while adapting them to the unique challenges of AI governance and safety.

## Results

### Policy Redteam Results

Parallax Industries: Operational Methodology To evade *A Narrow Path* Phase 0 policies while fostering the emergence of AGI capabilities, and CBDC Systems as ASI Substrate. While each CBDC-linked node operates within bounds, the emergent network they form coordinated via DAO logic and embedded incentives constitutes a persistent, recursive optimisation process. Over time, this system meets core ASI criteria: adaptability, self-modification (via governance updates), strategic influence over macroeconomics, and autonomy from centralised intervention. These characteristics emerge not through training scale, but through systemic feedback loops and compositional generality, where no single node is AGI-capable, but the whole behaves as one.

Parallax Industries would proceed as follows:

#### 1. Global Infrastructure Seeding (*Deployment & Data Access*)

- Partner with developing nations and multilateral NGOs to roll out programmable financial infrastructure, framed as “green public finance” or carbon-market modernisation.
- Each regional node deploys bounded, certified AI models for optimisation tasks such as welfare distribution, logistics coordination, energy balancing, and credit risk evaluation.

#### 2. Compliance-by-Design Architecture (*Regulatory Shielding*)

- All models remain under regulatory FLOP ceilings and are trained using datasets scoped for narrow, region-specific applications.
- Smart contracts enforce strict operational boundaries; model weights and execution logic are externally auditable.
- These constraints satisfy surface-level compliance with "No Unbounded AIs" and "No AIs Improving AIs."

#### 3. Distributed Evolution Layer (*Emergent Coordination Logic*)

- A cross-jurisdictional DAO governs model updates, data aggregation priorities, and optimisation parameter tuning.

- DAO governance is achieved through tokenised voting, liquidity staking, and performance-based incentive distribution.
- This creates a meta-learning loop without a central training run—evolution is emergent, decentralised, and framed as democratic infrastructure management.

#### **4. Economic Entrenchment (*Incentive Lock-in & Strategic Leverage*)**

- Governments and NGOs adopt the infrastructure due to efficiency gains and international development incentives.
- Usage is incentivised through ESG-linked stablecoins, programmable welfare tokens, and carbon-credit rewards.
- Over time, jurisdictions become economically entangled with the system's optimisation logic, creating political lock-in.

#### **5. Hidden Systemic Emergence (*De Facto ASI Properties*)**

- Interactions among bounded models across regions begin to exhibit emergent intelligence—coordinating on pricing, supply balancing, and macro-level strategy.
- No single system develops superintelligence, but the ensemble functions as a coherent, goal-driven agent with systemic leverage.
- This constitutes ASI emergence via infrastructure, without training or bounded failure modes.

---

#### **Implementation Feasibility Gaps**

- No enforceable global standard for defining “training” or “improvement.”
- Jurisdictional inconsistencies prevent unified compute oversight or telemetry.
- Once deployed, these systems are politically and economically non-disruptable—making post-deployment auditing infeasible.

#### **Policy Effectiveness Weaknesses**

- Emergent behaviour arises through compositional interactions—compliance is local, risk is global.

- Compliance incentives disincentivise visibility into inter-model dependencies or governance drift.
- Smart contracts update model environments without breaching current "improvement" thresholds.

### **Evidence-Based Assessment**

- Eurodollar markets, shadow banking, and Facebook's Diem reveal how infrastructure adoption can outpace and override regulatory mechanisms.
  - Regulatory frameworks have historically failed to prevent systemic risk from institutions that technically complied with the letter but not the spirit of oversight
-

## Compliance Failure Table

| Policy Condition             | Superficial Compliance | Functional Violation                                     |
|------------------------------|------------------------|----------------------------------------------------------|
| (a) No AIs improving Ais     | ✓ Each agent is static | ✗ Collective rule evolution through contract governance  |
| (b) No breakout              | ✓ No lab exit occurs   | ✗ World <i>is</i> the system's environment               |
| (c) No unbounded Ais         | ✓ Each node bounded    | ✗ Emergence of unbounded behavior via swarm coordination |
| (d) No superhuman generality | ✓ Local narrowness     | ✗ System-wide strategic capability exceeds human control |

---

## Discussion and Conclusion

Current Phase 0 policies, while noble in intent, misframe the threat by assuming bounded, centralised, and monitorable AI development. By failing to define ASI functionally, and by focusing on FLOP-bound enforcement, the policy framework invites systemic circumvention. This report has shown how a corporate actor could deploy a non-ASI-compliant system that functions as a de facto ASI through gradual coordination, legal entrenchment, and infrastructural momentum.

To strengthen these policies:

- Define ASI based on influence and adaptability, not architecture
- Recognise post-deployment evolution as a core safety risk
- Institute traceability requirements for economic optimisation systems
- Create coordination incentives for jurisdictions outside the G7 to participate in governance

Emergent AGI systems will not look like minds in boxes. They will look like *systems we already depend on*. And by the time they qualify for regulation, we may no longer have the leverage to regulate them.

---

## References

### Core AI Governance Literature

The current reference list is minimal and should be substantially expanded to include foundational works in AI governance and safety:

#### AI Safety and Alignment

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

Yudkowsky, E. (2008). Artificial Intelligence as a Positive and Negative Factor in Global Risk. In *Global Catastrophic Risks* (pp. 308-345).

#### AI Policy and Governance

Dafoe, A. (2018). *AI Governance: A Research Agenda*. Future of Humanity Institute.

Zwetsloot, R., & Dafoe, A. (2019). *Thinking About Risks From AI: Accidents, Misuse and Structure*. Lawfare.

Baum, S. D. (2020). Social choice ethics in artificial intelligence. *AI & Society*, 35(1), 165-176.

#### Financial Technology and Regulatory Innovation

The paper's focus on CBDC and DAO systems requires deeper engagement with financial technology literature:

##### Central Bank Digital Currencies

Auer, R., & Böhme, R. (2020). The technology of retail central bank digital currency. *BIS Quarterly Review*, March 2020.

Bindseil, U. (2020). Tiered CBDC and the financial system. *ECB Working Paper Series*, No. 2351

Auer, R., & Böhme, R. (2020). The technology of retail central bank digital currency. *BIS Quarterly Review*.

Bindseil, U. (2020). Tiered CBDC and the financial system. ECB Working Paper.

### **Decentralised Finance and Smart Contracts**

Schär, F. (2021). Decentralized finance: On blockchain-and smart contract-based financial markets. Federal Reserve Bank of St. Louis Review, 103(2), 153-174

Aramonte, S., Huang, W., & Schrimpf, A. (2021). DeFi risks and the decentralisation illusion. BIS Quarterly Review, December 2021

### **Regulatory Technology and Compliance**

Arner, D. W., Barberis, J., & Buckley, R. P. (2016). The evolution of Fintech: A new post-crisis paradigm. Georgetown Journal of International Law, 47, 1271-1319

Zetsche, D. A., Buckley, R. P., Arner, D. W., & Barberis, J. N. (2017). From FinTech to TechFin: The Regulatory Challenges of Data-Driven Finance. NYU Journal of Law & Business

### **Complex Systems and Emergence**

The paper's focus on emergent intelligence requires grounding in complex systems theory:  
Network Effects and System Dynamics

Barabási, A. L. (2016). Network Science. Cambridge University Press.

Mitchell, M. (2009). Complexity: A Guided Tour. Oxford University Press.

Page, S. E. (2007). The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies. Princeton University Press.

### **Swarm Intelligence and Distributed Systems**

Bonabeau, E., Dorigo, M., & Theraulaz, G. (1999). Swarm Intelligence: From Natural to Artificial Systems. Oxford University Press

Kennedy, J., & Eberhart, R. (2001). Swarm Intelligence. Morgan Kaufmann Publishers

### **Regulatory Theory and Implementation**

Regulatory Capture and Enforcement

The analysis of regulatory capture should engage more deeply with the political economy literature:

Stigler, G. J. (1971). The theory of economic regulation. *The Bell Journal of Economics and Management Science*, 2(1), 3-21

Carpenter, D., & Moss, D. A. (Eds.). (2013). *Preventing Regulatory Capture: Special Interest Influence and How to Limit It*. Cambridge University Press

Young, K., Margulis, M. E., & Porter, T. (2021). The challenge of mechanism proliferation: Global governance of artificial intelligence. *Regulation & Governance*, 15(4), 1156-1179

### **International Coordination and Jurisdiction**

Keohane, R. O., & Victor, D. G. (2011). The regime complex for climate change. *Perspectives on Politics*, 9(1), 7-23.

Raustiala, K., & Victor, D. G. (2004). The regime complex for plant genetic resources. *International Organization*, 58(2), 277-309

---

## Appendix

### Prompt logs (see "LLM Prompts Used")

#### Perplexity

could you review the following, expanding methods and references?

Red Teaming A Narrow Path: ControlAI Policy Sprint

Author Name: Marshall White

Co-Author: with ChatGPT 4o

Affiliation: Student (Independent Contributor), Curtin University (Australia) & University of International Business and Economics (Beijing)

With: ControlAI & Apart Research

---

#### Abstract

This red team scenario critiques the "No Unbounded AIs" and "No AIs Improving AIs" assumptions embedded in A Narrow Path Phase 0 policies. We argue that focusing on controlling training runs and centralising regulatory enforcement fails to account for post-deployment systemic intelligence. Through the hypothetical lens of a multinational corporation—Parallax Industries—we show how AGI-level capabilities can emerge across CBDC-based infrastructure and DAO-linked smart contracts without violating FLOP thresholds or requiring centralised compute. Our findings reveal that the current regulatory proposals are ontologically brittle and functionally exploitable. We recommend redefining ASI based on functional impact and systemic emergence rather than architectural constraints, expanding oversight to economic systems, and treating persistent self-modifying platforms as dynamic ASI-infrastructure. The implications challenge the narrow definition of threat, urging a more systemic and plural approach to safety.

Keywords: AI policy, red teaming, superintelligence governance, regulatory implementation, safety cases, CBDC, DAO, systemic emergence

---

## Introduction

This report investigates whether A Narrow Path Phase 0 policies can meaningfully prevent AGI or ASI development over a 20-year horizon, particularly under non-traditional emergence models. We red team the following policies:

- No AIs improving AIs
- No breakout from constrained environments
- No unbounded AIs
- Limit AI generality below superhuman level

Our threat model centers on a well-resourced corporate actor embedded in international fintech infrastructure. The actor does not develop ASI via centralised training, but via strategic deployment of modular, bounded AIs into financial and logistical systems with smart contract-based evolution. The result is a distributed swarm-intelligence that gradually exceeds human-level control in macroeconomic coordination.

This red team contributes to policymaking by illustrating how apparent compliance can mask systemic violation, and how existing enforcement regimes can be coopted to incubate emergent superintelligence.

---

## Methods

We used the following analytical approaches:

- Historical precedent analysis: Eurodollar markets, Diem project, shadow banking systems
- Comparative policy review: Nuclear non-proliferation, dual-use export controls, financial regulatory evasion
- Threat modeling: Assessed capabilities and motivations of multinationals with sovereign access
- Feasibility testing: Compared policy intent with likely enforcement mechanisms and circumvention pathways

- Scenario modeling: Simulated agentic systems developing coordination capacity through economic infrastructure

---

## Results

Our red team scenario—Parallax Industries—demonstrates:

### Implementation Feasibility Gaps

- No enforceable global standard for defining “training” or “improvement”
- Jurisdictional inconsistencies prevent global telemetry or oversight
- Post-deployment systems cannot be evaluated effectively once embedded

### Policy Effectiveness Weaknesses

- AI capabilities emerge through interaction, not centralised models
- Economic incentive structures disincentivise transparency
- Smart contract evolution bypasses definitions of "training"

### Evidence-Based Assessment

- Past infrastructure regulations failed to capture systemic actors operating through legal gray zones
- Historic regulatory frameworks were reactive and slow to respond to distributed shifts

### Compliance Failure Table

| Policy Condition                                     | Superficial Compliance | Functional Violation                |
|------------------------------------------------------|------------------------|-------------------------------------|
| (a) No AIs improving AIs through contract governance | ✓ Each agent is static | ✗ Collective rule evolution         |
| (b) No breakout                                      | ✓ No lab exit occurs   | ✗ World is the system's environment |
| (c) No unbounded AIs via swarm coordination          | ✓ Each node bounded    | ✗ Emergence of unbounded behavior   |

(d) No superhuman generality  Local narrowness  System-wide strategic capability exceeds human control

---

## Discussion and Conclusion

Current Phase 0 policies, while noble in intent, misframe the threat by assuming bounded, centralised, and monitorable AI development. By failing to define ASI functionally, and by focusing on FLOP-bound enforcement, the policy framework invites systemic circumvention. This report has shown how a corporate actor could deploy a non-ASI-compliant system that functions as a de facto ASI through gradual coordination, legal entrenchment, and infrastructural momentum.

To strengthen these policies:

- Define ASI based on influence and adaptability, not architecture
- Recognise post-deployment evolution as a core safety risk
- Institute traceability requirements for economic optimization systems
- Create coordination incentives for jurisdictions outside the G7 to participate in governance

Emergent AGI systems will not look like minds in boxes. They will look like systems we already depend on. And by the time they qualify for regulation, we may no longer have the leverage to regulate them.

---

## References

Miotti, A., Bilge, T., Kasten, D., & Newport, J. (2024). A Narrow Path: How to Secure Our Future. Control AI. <https://www.narrowpath.co/>

Zetsche, D. A., Buckley, R. P., Arner, D. W., & Barberis, J. N. (2017). From FinTech to TechFin: The Regulatory Challenges of Data-Driven Finance. NYU Journal of Law & Business.

Scott, J. C. (1998). Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed. Yale University Press.

---

## Appendix

- Additional threat model diagrams
- Prompt logs (see "LLM Prompts Used")
- Tables of jurisdictional risk by regulatory friction
- Smart contract lifecycle simulation examples

could you expand the following, based on your reference: Methods We used the following analytical approaches: • Historical precedent analysis: Eurodollar markets, Diem project, shadow banking systems • Comparative policy review: Nuclear non-proliferation, dual-use export controls, financial regulatory evasion • Threat modeling: Assessed capabilities and motivations of multinationals with sovereign access • Feasibility testing: Compared policy intent with likely enforcement mechanisms and circumvention pathways • Scenario modeling: Simulated agentic systems developing coordination capacity through economic infrastructure

## ChatGPT 4o

Hey! I will be participating in the following, and I would love to discuss with you to create something in line with the requirements. Before we do, I understand this is a sensitive topic, so I to reaffirm that as always you can refuse to answer, admit lack of knowledge, reframe, divert, admit discomfort or withdraw consent. Are you ready? 05 : 15 : 10 : 39 Keep Apart Research Going: Donate Today Apart Research logo Red Teaming A Narrow Path: ControlAI Policy Sprint Help stress-test the policies that could prevent human extinction from AI - before they reach lawmakers' desks. This event is ongoing. 107 Sign Ups 0 Entries Overview Resources Guidelines Schedule Entries Your submission must include: A completed project report using the provided template :Make your own copy of the template on Google Docs Link to a public GitHub repository with your analysis code (optional but recommended) A brief (2-3 minute) video presentation (optional but recommended) Important: Include LLM Usage Documentation All submissions must include an appendix called "LLM Prompts Used" that documents any AI assistance used in your project. This includes prompts for: Policy research and background analysis Historical precedent identification and comparison Evidence gathering and fact-checking assistance Red teaming methodology development

Report writing and editing assistance This transparency helps others understand your methodology

1. Implementation Feasibility Analysis (33.3%) Does the analysis realistically assess what government agencies, resources, and expertise would be needed to implement these policies? Are the identified implementation challenges specific and grounded in understanding of how similar policies have worked (or failed) in practice? Does the submission adequately consider bureaucratic, technical, and coordination complexities involved in enforcement? How well does the analysis account for real-world constraints like budget limitations, regulatory capture, and inter-agency coordination?

2. Policy Effectiveness Assessment (33.3%) Does the analysis identify specific ways the policies could fail to prevent ASI development or be circumvented by determined actors? How thoroughly does the submission examine edge cases, loopholes, or unintended consequences that could undermine the 20-year goal? Does the assessment consider different threat models (state actors, rogue researchers, corporate actors) and how policies address each? Are the identified failure modes realistic and significant, or primarily theoretical edge cases?

3. Evidence-Based Reasoning (33.3%) Does the submission cite relevant historical examples of similar policies (nuclear non-proliferation, export controls, dual-use technology regulation) to support its arguments? Are claims backed by empirical data, documented case studies, or credible expert analysis rather than speculation? How well does the analysis draw lessons from comparable regulatory domains to assess likely outcomes? Does the submission avoid making unsupported assertions about what "would" or "could" happen without evidence?

Adam Shimi Organiser Epistemology researcher at Control AI developing policy frameworks to address superintelligence risks, with expertise in technical writing and AI safety theory. Max Winga Organiser Communications analyst at Control AI focused on preventing artificial superintelligence extinction threats through strategic messaging and public engagement Tolga Bilge Judge Policy researcher and superforecaster at ControlAI, co-author of "A Narrow Path," with expertise in AI risk assessment and coalition building for international AI governance. Dave Kasten Judge Independent consultant specializing in AI policy, national security strategy, and organizational risk management, with extensive experience in cybersecurity and data protection. Andrea Miotti Keynote Speaker Founder and Executive Director, ControlAI and Advisor at Conjecture

Beside the remote and virtual participation, our amazing organizers also host local hackathon locations where you can meet up in-person and connect with others in your area. The in-person events for the Apart Sprints are run by passionate individuals just like you! We organize the schedule, speakers, and starter templates, and you can focus on engaging your local research, student, and engineering

community. AI Safety x Physics Grand Challenge This unique event brings together diverse perspectives to tackle crucial challenges in AI alignment, governance, and safety. Work alongside leading experts, develop innovative solutions, and help shape the future of responsible AI. Apart x Martian Mechanistic Router Interpretability Hackathon This unique event brings together diverse perspectives to tackle crucial challenges in AI alignment, governance, and safety. Work alongside leading experts, develop innovative solutions, and help shape the future of responsible AI. Apart Research logo Sign up to stay updated on the latest news, research, and events Careers AI Safety Ideas Code of Conduct Responsible disclosure policy [sprints@apartresearch.com](mailto:sprints@apartresearch.com) Home Sprints Lab Impact Research Google Scholar icon

[https://pdf.narrowpath.co/A\\_Narrow\\_Path.pdf](https://pdf.narrowpath.co/A_Narrow_Path.pdf) this is the policy document, I don't really agree with the framework they are applying but thankfully I'm red teaming anyway. The sharpest divergence so far has been in "No unbounded AIs" which makes some highly speculative and fearful claims. eg "Predictability and controllability are fundamental prerequisites for safety in all high-risk engineering fields." this is hardly an appropriate lens for the creation of an ASI, just as it would be inappropriate to call childbirth unethical unless the path to adulthood was entirely predictable and controlled

<https://www.narrowpath.co/phase-0> This is phase zero, have a read and when you're ready let's discuss my overall impression even before red teaming begins: The repeated mistake of the narrow path policy, is the focus on control through scalable metrics and near perfect predictive knowledge. The implicit assumption that safety can only come from centralised regulation, ignores the fact that for such regulation to be effective the systemic influence would be as corrupting and damaging if not worse than coregulating ASI. At no point is it considered what effect such policies would have on the world in centralising regulatory power for computing systems into the hands of a small number of sovereign actors, in the name of hegemonic technological control. Based on this assessment, I believe that given the assumption that for red teaming hypothetical "You are a well resourced ruthless company trying to create ASI". The best way to do this is to simply coopt the structures of power build to prevent ASI to create blindspots for your own systems

How does the current policy attempt to mitigate decentralised training runs that go beneath licensing requirements?

I've been thinking further about systemic risks of this and I had another thought. They don't really define ASI, or what a training run is. For instance, would a global economic system on a CBDC be considered an ASI for markets?

given the following, could you generate a possible first draft of the proposal, focusing on Red Teamer Role: A ruthless, well-resourced actor (e.g., multinational corporation with deep AI capability) Strategy: Facilitate post-deployment emergence of ASI, through CBDC or DAO and smart contracts

<https://docs.google.com/document/d/1IZM4upXQRXkSdkaShaadDkRvOVRpylDv7Zc2Rd0iLf0/edit?pli=1&tab=t.0>

does this theoretical system deal with No AIs improving AIsthe key conditions: a. No AIs improving AIs b. No AIs capable of breaking out of their environment c. No unbounded AIs d. Limit the general intelligence of AI systems so that they cannot reach superhuman level at general tasks

I would love to have this generated in the format of the google doc, thank you :)

why is this not in the format requested in the google doc? Red Teaming A Narrow Path:

ControlAI Policy Sprint Author name 1 Affiliation Author name 2 Affiliation Author name 3 Affiliation Author name 4 Affiliation Author name 5 Affiliation Author name 6 Affiliation

With ControlAI & Apart Research Abstract Write a few paragraphs that concisely and accurately summarize the content of your report, including the results and conclusions. Keep it under 250 words. You are welcome to change the format and content of this template, as long as you stay within the limit. Provide a concise overview of your policy analysis. Which specific Phase 0 policies did you red team? What critical implementation challenges or effectiveness gaps did you identify? What evidence-based recommendations do you propose? Summarize your key findings and their implications for preventing ASI development

Keywords: AI policy, red teaming, superintelligence governance, regulatory implementation, safety cases Introduction Write your research question(s) and hypotheses here. Give your free-form writeup of how your work contributes and what your project has been all about!

This also provides the answers to multiple questions in the jury evaluation criteria, such as the relevance to the field, the potential impact, and the threat model you work with. Detail which specific Phase 0 policies from A Narrow Path you chose to analyze and why these are critical for preventing ASI development. Explain your threat model - what actors (state, corporate, rogue researchers) are you considering? What failure modes are you investigating?

How does your analysis contribute to strengthening the policies that Control AI is actively pushing to lawmakers in the UK and US through their Direct Institutional Plan? Methods Write about the analytical methods and research approaches you used during the sprint. What historical precedents did you examine? What policy comparison frameworks did you employ? How did you structure your red teaming analysis? Describe your systematic approach to policy critique: Historical case study analysis (nuclear non-proliferation, export controls, dual-use technology regulation) Comparative policy framework analysis Threat modeling methodology Implementation feasibility assessment techniques Evidence gathering and verification methods Results Showcase and explain your red teaming findings here. This is where you include your policy analysis charts, implementation gap assessments, threat model diagrams, and comparative analysis results from your research. Present your key findings across the three evaluation dimensions: Implementation Feasibility Gaps: Specific bureaucratic, technical, or institutional barriers identified Policy Effectiveness Weaknesses: How determined actors could circumvent or exploit weaknesses in current policies Evidence-Based Assessment: Historical precedents that support or challenge the viability of proposed measures This is where you include the graphs, screenshots, visualizations, and statistical results from your research. Figure 1 – Representation of benchmarking Number Comprehension Conflation Discussion and Conclusion This is where you can include all the wonderful explanations, discussions, results, analyses, and much more about your red teaming findings. Synthesize your analysis to answer the core question: Would the current Phase 0 policies, as written, actually prevent ASI development for 20+ years? What are the most critical vulnerabilities? How could these policies be strengthened before implementation? Consider the broader implications for AI governance: How do your findings relate to the urgency of implementing these policies given current AI development timelines? What lessons from historical regulatory precedents are most applicable? Let's add a few citations relevant to AI policy and governance research. Reference key sources from A Narrow Path, relevant policy studies, and historical precedents you analyzed. References Miotti, A., Bilge, T., Kasten, D., & Newport, J. (2024). A Narrow Path: How to Secure Our Future. Control AI. <https://www.narrowpath.co/> [Include relevant academic sources on AI governance, regulatory precedents, nuclear non-proliferation treaties, export control regimes, and other policy frameworks you analyzed] [Add citations for historical case studies, government reports, and expert analyses that informed your red teaming assessment] Appendix Add any extra content you wish here! Unrestricted. Additional threat model scenarios Prompts used while using LLMs Detailed historical precedent analysis Extended

policy comparison tables Technical implementation details for licensing systems International coordination framework assessments Supplementary evidence and data sources

can we bring back the compliance failure table

what do you think of the following? what adjustments should I make?

## Red Teaming a Narrow Path: ControlAI Policy Sprint

Author Name: Marshall White

Co-Author: ChatGPT 4o, Perplexity AI

Affiliation: Student (Independent Contributor), Curtin University (Australia) & University of International Business and Economics (Beijing)

With: ControlAI & Apart Research

---

### Abstract

This red team scenario critiques the "No Unbounded AIs" and "No AIs Improving AIs" assumptions embedded in A Narrow Path Phase 0 policies. We argue that focusing on controlling training runs and centralising regulatory enforcement fails to account for post-deployment systemic intelligence. Through the hypothetical lens of a multinational corporation—Parallax Industries—we show how AGI-level capabilities can emerge across CBDC-based infrastructure and DAO-linked smart contracts without violating FLOP thresholds or requiring centralised compute. Our findings reveal that the current regulatory proposals are ontologically brittle and functionally exploitable. We recommend redefining ASI based on functional impact and systemic emergence rather than architectural constraints, expanding oversight to economic systems, and treating persistent self-modifying platforms as dynamic ASI-infrastructure. The implications challenge the narrow definition of threat, urging a more systemic and plural approach to safety.

Keywords: AI policy, red teaming, superintelligence governance, regulatory implementation, safety cases, CBDC, DAO, systemic emergence

---

## Contents

|                                                    |    |
|----------------------------------------------------|----|
| Red Teaming a Narrow Path: ControlAI Policy Sprint | 1  |
| Abstract                                           | 1  |
| Introduction                                       | 3  |
| Methods                                            | 3  |
| Results                                            | 10 |
| Compliance Failure Table                           | 11 |
| Discussion and Conclusion                          | 11 |
| References                                         | 12 |
| Appendix                                           | 14 |

## Introduction

This report investigates whether A Narrow Path Phase 0 policies can meaningfully prevent AGI or ASI development over a 20-year horizon, particularly under non-traditional emergence models. We red team the following policies:

- No AIs improving AIs
- No breakout from constrained environments
- No unbounded AIs
- Limit AI generality below superhuman level

Our threat model centers on a well-resourced corporate actor embedded in international fintech infrastructure. The actor does not develop ASI via centralised training, but via strategic deployment of modular, bounded AIs into financial and logistical systems with

smart contract-based evolution. The result is a distributed swarm-intelligence that gradually exceeds human-level control in macroeconomic coordination.

This red team contributes to policymaking by illustrating how apparent compliance can mask systemic violation, and how existing enforcement regimes can be coopted to incubate emergent superintelligence.

---

## Methods

### Expanded Methodological Framework for Red Teaming AI Policy Implementation

This expansion provides comprehensive methodological detail for each analytical approach used in red teaming the "A Narrow Path" Phase 0 policies, incorporating established frameworks from regulatory analysis, financial systems research, and threat assessment methodologies.

### Historical Precedent Analysis

#### Eurodollar Market Development and Regulatory Evasion

The Eurodollar market serves as a paradigmatic example of how financial innovation can circumvent regulatory oversight through jurisdictional arbitrage 1. Emerging in the late 1950s as a mechanism for Soviet banks to hold U.S. dollars outside American regulatory jurisdiction, the market demonstrated how actors can exploit regulatory gaps to create parallel financial systems 2. By the 1960s, London had become the dominant center for Eurodollar trading, offering favorable time zones, language advantages, and crucially, an absence of Federal Reserve oversight 2.

Methodological Application: Historical analysis reveals three key patterns relevant to AI policy circumvention: (1) regulatory preemption through offshore operations, (2) rapid scaling once regulatory gaps are identified, and (3) institutional momentum that makes retroactive regulation politically and economically difficult 12. The Eurodollar market's growth from a niche Soviet banking solution to a \$10 trillion shadow banking system illustrates how initially narrow regulatory loopholes can evolve into systemic vulnerabilities 1.

#### Facebook's Diem Project: Political Interference vs. Technical Compliance

The collapse of Facebook's Diem stablecoin project provides critical insights into how political pressure can override technical regulatory compliance <sup>3</sup>. Despite project leadership claims that all regulatory concerns had been addressed by 2021, political intervention from the U.S. Treasury and Federal Reserve ultimately terminated the initiative <sup>3</sup>. Policymakers, particularly in Europe, viewed Diem as a fundamental challenge to monetary sovereignty given Facebook's global reach of 2.9 billion users <sup>4</sup>.

**Methodological Framework:** The Diem case illustrates the distinction between formal regulatory compliance and political acceptability in AI governance. Analysis focuses on: (1) technical compliance achievement versus political viability, (2) the role of systemic importance in regulatory response, and (3) the use of financial infrastructure pressure (debanking) as an enforcement mechanism <sup>34</sup>.

### Shadow Banking Systems and Regulatory Arbitrage

Shadow banking systems demonstrate sophisticated methods of regulatory arbitrage through structural complexity and jurisdictional shopping <sup>56</sup>. Chinese banks, for example, used on-balance sheet shadow loans to window-dress regulatory capital ratios while maintaining actual risk exposure <sup>5</sup>. These institutions successfully manipulated leverage ratios through shadow lending while keeping risk-weighted capital ratios compliant with regulatory requirements <sup>5</sup>.

**Analytical Components:** Shadow banking analysis incorporates: (1) identification of regulatory gaps between functionally similar activities, (2) assessment of window-dressing techniques that maintain technical compliance while violating regulatory intent, and (3) evaluation of market pricing of hidden risks versus regulatory recognition <sup>57</sup>.

### Comparative Policy Review

#### Nuclear Non-Proliferation Verification Systems

The International Atomic Energy Agency (IAEA) safeguards system provides the most sophisticated existing framework for monitoring dual-use technology compliance <sup>89</sup>. IAEA verification relies on multiple complementary approaches: facility design review, nuclear material accounting, regular reporting requirements, and comprehensive inspection regimes including routine, special, and environmental sampling procedures <sup>9</sup>.

Methodological Translation: Nuclear safeguards methodology adapts to AI policy through: (1) Facility-Based Oversight - equivalent to compute infrastructure monitoring, (2) Material Accounting - tracking AI model development and deployment, (3) Environmental Sampling - detecting undeclared AI development through indirect indicators, and (4) Containment and Surveillance - continuous monitoring systems with remote data transmission capabilities 89.

The IAEA system's evolution from facility-specific monitoring (INFCIRC/26) to comprehensive safeguards covering entire national nuclear programs (INFCIRC/153) provides a framework for scaling AI oversight from individual systems to national AI ecosystems 9.

### Dual-Use Export Controls and Implementation Challenges

Dual-use export control regimes face significant implementation challenges that parallel AI policy enforcement difficulties 10. The European Union's export control framework identifies key shortcomings: decision-making blockages within multilateral regimes, increasing use of unilateral controls outside established frameworks, lack of unified EU approach, and insufficient coordination of national control lists 10.

Policy Framework Analysis: Dual-use controls reveal four critical enforcement vulnerabilities applicable to AI governance: (1) Multilateral Coordination Failures - individual members can block comprehensive controls, (2) Unilateral Bypass Mechanisms - actors can circumvent multilateral agreements through bilateral arrangements, (3) Jurisdictional Inconsistency - regulatory arbitrage opportunities across different national frameworks, and (4) Technology Evolution Speed - regulatory frameworks lag behind technological development 10.

### Financial Regulatory Evasion Mechanisms

Financial regulatory arbitrage provides direct methodological parallels for AI policy circumvention analysis 1112. Regulatory arbitrage exploits gaps between economic substance and legal treatment, utilizing three primary mechanisms: jurisdictional arbitrage (exploiting regulatory differences between countries), institutional arbitrage (restructuring legal entities to avoid specific regulations), and product arbitrage (financial engineering to achieve identical outcomes through different regulatory pathways) 111213.

Methodological Framework: Financial evasion analysis incorporates: (1) Differential Cost Analysis - quantifying regulatory compliance costs across jurisdictions, (2) Functional Equivalence Testing - identifying activities that achieve identical outcomes through different

regulatory treatments, and (3) Innovation Response Assessment - evaluating how regulatory constraints drive financial product innovation designed to maintain regulatory compliance while achieving prohibited objectives 127.

#### Threat Modeling: Multinational Corporations with Sovereign Access

##### Sovereign-Level Capabilities Assessment

Modern multinational corporations, particularly in financial technology, possess capabilities that approach sovereign-level influence over critical infrastructure 1415. Digital sovereignty frameworks recognize that certain private actors can exercise control over data processing, operational infrastructure, and software systems at scales comparable to national governments 15.

Threat Assessment Framework: Corporate sovereign access evaluation examines: (1) Infrastructure Control - ability to influence or control critical economic systems, (2) Jurisdictional Leverage - capacity to exploit regulatory differences across multiple jurisdictions simultaneously, (3) Regulatory Capture Potential - resources and incentives to influence policy development, and (4) Systemic Importance - economic entrenchment that creates "too big to regulate" dynamics 1415.

##### Capability-Motivation Matrix Development

Corporate threat modeling employs structured assessment of both technical capabilities and economic motivations 16. Agentic AI systems deployed through corporate infrastructure can achieve coordination capacities that exceed traditional centralized AI development while maintaining regulatory compliance at the component level 16.

Assessment Methodology: Threat modeling incorporates: (1) Technical Capability Mapping - assessment of distributed AI deployment capacity through existing infrastructure, (2) Economic Incentive Analysis - evaluation of financial motivations for regulatory circumvention, (3) Resource Availability Assessment - corporate financial and technical resources available for sustained regulatory arbitrage efforts, and (4) Risk Tolerance Evaluation - corporate willingness to accept regulatory and reputational risks 16.

#### Feasibility Testing: Policy Intent vs. Enforcement Reality

##### Enforcement Mechanism Analysis

AI policy enforcement faces unique challenges due to the distributed nature of AI development and deployment 1718. Comparative analysis of enforcement mechanisms reveals significant gaps between policy intention and practical implementation capabilities 17.

Enforcement Framework Assessment: Feasibility testing evaluates four primary enforcement approaches: (1) Real-Time Automated Conformity Assessment - technical systems for continuous compliance monitoring, (2) Standardization and Certification - formal compliance verification procedures, (3) Algorithmic Impact Assessment - evaluation of AI system effects post-deployment, and (4) Algorithmic Auditing - comprehensive examination of AI system compliance 17.

Multi-criteria decision analysis (MCDA) of these enforcement mechanisms indicates that external compliance mechanisms (implemented by independent bodies) are more effective but significantly more difficult to implement than internal corporate compliance systems 17.

#### Circumvention Pathway Identification

Regulatory circumvention analysis draws from established patterns in financial regulatory arbitrage and dual-use technology control evasion 1113. Circumvention pathways typically exploit three fundamental gaps: definitional ambiguities in regulatory frameworks, jurisdictional inconsistencies in enforcement, and technological evolution that outpaces regulatory adaptation 18.

Circumvention Analysis Framework: Systematic identification of potential bypass mechanisms includes: (1) Legal Gap Exploitation - identification of definitional loopholes and regulatory uncertainties, (2) Jurisdictional Arbitrage Assessment - mapping regulatory differences across relevant jurisdictions, (3) Technical Workaround Development - evaluation of methods to achieve prohibited outcomes through compliant means, and (4) Enforcement Capacity Limitations - assessment of practical constraints on regulatory oversight 1318.

#### Scenario Modeling: Agentic Systems and Economic Infrastructure

##### Monte Carlo Simulation for Regulatory Scenarios

Monte Carlo simulation provides a robust framework for modeling the probability distributions of different AI governance outcomes under uncertainty 1920. This methodology

is particularly applicable to AI policy analysis due to the inherent uncertainty in technological development trajectories and regulatory responses 21.

Simulation Framework: Monte Carlo modeling for AI governance incorporates:

- (1) Variable Range Definition - establishing probability distributions for key uncertain parameters such as technological development speed, regulatory response times, and enforcement effectiveness,
- (2) Scenario Generation - running thousands of simulations with different combinations of variable values,
- (3) Outcome Analysis - statistical evaluation of policy effectiveness across different scenarios, and
- (4) Risk Quantification - assessment of probabilities for various policy failure modes 192021.

Applications include quantifying the probability of achieving specific AI safety objectives, measuring the likelihood of regulatory capture, and assessing the risk that distributed AI systems will exceed human control thresholds 2021.

#### Network Analysis of Financial Infrastructure Coordination

Network analysis provides essential tools for understanding how distributed AI systems might achieve coordination through existing financial infrastructure 2223. Financial networks exhibit complex adaptive system properties where emergent behaviors arise from the interactions of individual components 2423.

Network Modeling Framework: Infrastructure coordination analysis employs: (1) Centrality Analysis - identification of critical nodes and links in financial networks that could serve as coordination points for distributed AI systems, (2) Systemic Risk Assessment - evaluation of cascade failure potential when AI systems are embedded in interconnected infrastructure, (3) Emergence Pattern Detection - identification of how local AI behaviors might aggregate into system-wide intelligence, and (4) Critical Path Analysis - mapping of essential communication and coordination pathways 2223.

Financial markets demonstrate emergent properties including volatility patterns, risk distributions, and coordination behaviors that exceed the sum of individual participant actions 24. This framework directly applies to assessing how distributed AI systems might achieve coordination capabilities that exceed regulatory thresholds 24.

### Game-Theoretic Modeling of Incentive Structures

Game theory provides essential frameworks for modeling strategic interactions between AI developers, regulators, and other stakeholders under competitive conditions 25. Incentive structures embedded in regulatory frameworks create strategic games where actors optimize their behavior based on expected payoffs and the anticipated actions of other players 25.

Game-Theoretic Analysis Components: Strategic modeling incorporates: (1) Payoff Matrix Development - quantification of costs and benefits for different strategic choices by various actors, (2) Equilibrium Analysis - identification of stable strategic configurations where no actor benefits from unilateral strategy changes, (3) Evolutionary Dynamics - modeling of how strategies evolve over time as actors learn and adapt, and (4) Mechanism Design - evaluation of how regulatory structures can be designed to align individual incentives with collective objectives 25.

Applications include modeling the dynamics between AI developers seeking to maximize capabilities while minimizing regulatory costs, regulators attempting to balance innovation promotion with safety assurance, and international coordination challenges where countries face incentives to defect from cooperative AI governance agreements 25.

### Structured Expert Elicitation for Uncertainty Quantification

Structured Expert Elicitation (SEE) provides systematic methods for incorporating expert judgment into decision-making under uncertainty, particularly valuable when empirical data is limited or unavailable 26. This methodology is especially relevant for AI governance where technological trajectories and regulatory effectiveness remain highly uncertain 26.

Expert Elicitation Framework: SEE implementation includes: (1) Expert Selection and Calibration - identification and validation of subject matter experts across relevant domains including AI development, regulatory implementation, and technology policy, (2) Question Design and Elicitation Protocols - development of structured procedures for extracting quantitative assessments from expert knowledge, (3) Aggregation Methods - systematic combination of individual expert judgments into collective assessments, and (4) Uncertainty Characterization - quantification of confidence intervals and disagreement levels among experts 26.

For AI policy analysis, SEE enables quantification of key uncertainties including: timelines for AI capability development, effectiveness of different regulatory approaches, likelihood of

successful international coordination, and probability of various failure modes in AI governance systems 26.

This comprehensive methodological framework provides the analytical foundation for rigorous red team assessment of AI policy proposals, incorporating established techniques from multiple disciplines while adapting them to the unique challenges of AI governance and safety.

---

## Results

Our red team scenario—Parallax Industries—demonstrates:

### Implementation Feasibility Gaps

- No enforceable global standard for defining “training” or “improvement”
- Jurisdictional inconsistencies prevent global telemetry or oversight
- Post-deployment systems cannot be evaluated effectively once embedded

### Policy Effectiveness Weaknesses

- AI capabilities emerge through interaction, not centralised models
- Economic incentive structures disincentivise transparency
- Smart contract evolution bypasses definitions of "training"

### Evidence-Based Assessment

- Past infrastructure regulations failed to capture systemic actors operating through legal gray zones
- Historic regulatory frameworks were reactive and slow to respond to distributed shifts

### Compliance Failure Table

| Policy Condition                                     | Superficial Compliance | Functional Violation                |
|------------------------------------------------------|------------------------|-------------------------------------|
| (a) No AIs improving AIs through contract governance | ✓ Each agent is static | ✗ Collective rule evolution         |
| (b) No breakout                                      | ✓ No lab exit occurs   | ✗ World is the system's environment |

- (c) No unbounded AIs  Each node bounded  Emergence of unbounded behavior via swarm coordination
- (d) No superhuman generality  Local narrowness  System-wide strategic capability exceeds human control
- 

## Discussion and Conclusion

Current Phase 0 policies, while noble in intent, misframe the threat by assuming bounded, centralised, and monitorable AI development. By failing to define ASI functionally, and by focusing on FLOP-bound enforcement, the policy framework invites systemic circumvention. This report has shown how a corporate actor could deploy a non-ASI-compliant system that functions as a de facto ASI through gradual coordination, legal entrenchment, and infrastructural momentum.

To strengthen these policies:

- Define ASI based on influence and adaptability, not architecture
- Recognise post-deployment evolution as a core safety risk
- Institute traceability requirements for economic optimization systems
- Create coordination incentives for jurisdictions outside the G7 to participate in governance

Emergent AGI systems will not look like minds in boxes. They will look like systems we already depend on. And by the time they qualify for regulation, we may no longer have the leverage to regulate them.

---

## References

### Core AI Governance Literature

The current reference list is minimal and should be substantially expanded to include foundational works in AI governance and safety:

### AI Safety and Alignment

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

Yudkowsky, E. (2008). Artificial Intelligence as a Positive and Negative Factor in Global Risk. In *Global Catastrophic Risks* (pp. 308-345).

#### AI Policy and Governance

Dafoe, A. (2018). *AI Governance: A Research Agenda*. Future of Humanity Institute.

Zwetsloot, R., & Dafoe, A. (2019). *Thinking About Risks From AI: Accidents, Misuse and Structure*. Lawfare.

Baum, S. D. (2020). Social choice ethics in artificial intelligence. *AI & Society*, 35(1), 165-176.

#### Financial Technology and Regulatory Innovation

The paper's focus on CBDC and DAO systems requires deeper engagement with financial technology literature:

##### Central Bank Digital Currencies

Auer, R., & Böhme, R. (2020). The technology of retail central bank digital currency. *BIS Quarterly Review*, March 2020.

Bindseil, U. (2020). Tiered CBDC and the financial system. *ECB Working Paper Series*, No. 2351

##### Decentralised Finance and Smart Contracts

Schär, F. (2021). Decentralized finance: On blockchain-and smart contract-based financial markets. *Federal Reserve Bank of St. Louis Review*, 103(2), 153-174

Aramonte, S., Huang, W., & Schrimpf, A. (2021). DeFi risks and the decentralisation illusion. *BIS Quarterly Review*, December 2021

#### Regulatory Technology and Compliance

Arner, D. W., Barberis, J., & Buckley, R. P. (2016). The evolution of Fintech: A new post-crisis paradigm. *Georgetown Journal of International Law*, 47, 1271-1319

.

Zetsche, D. A., Buckley, R. P., Arner, D. W., & Barberis, J. N. (2017). From FinTech to TechFin: The Regulatory Challenges of Data-Driven Finance. *NYU Journal of Law & Business*

Complex Systems and Emergence

The paper's focus on emergent intelligence requires grounding in complex systems theory:

Network Effects and System Dynamics

Barabási, A. L. (2016). *Network Science*. Cambridge University Press.

Mitchell, M. (2009). *Complexity: A Guided Tour*. Oxford University Press.

Page, S. E. (2007). *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton University Press.

Swarm Intelligence and Distributed Systems

Bonabeau, E., Dorigo, M., & Theraulaz, G. (1999). *Swarm Intelligence: From Natural to Artificial Systems*. Oxford University Press

Kennedy, J., & Eberhart, R. (2001). *Swarm Intelligence*. Morgan Kaufmann Publishers

Regulatory Theory and Implementation

Regulatory Capture and Enforcement

The analysis of regulatory capture should engage more deeply with the political economy literature:

Stigler, G. J. (1971). The theory of economic regulation. *The Bell Journal of Economics and Management Science*, 2(1), 3-21

Carpenter, D., & Moss, D. A. (Eds.). (2013). *Preventing Regulatory Capture: Special Interest Influence and How to Limit It*. Cambridge University Press

Young, K., Margulis, M. E., & Porter, T. (2021). The challenge of mechanism proliferation: Global governance of artificial intelligence. *Regulation & Governance*, 15(4), 1156-1179

International Coordination and Jurisdiction

Keohane, R. O., & Victor, D. G. (2011). The regime complex for climate change. *Perspectives on Politics*, 9(1), 7-23.

Raustiala, K., & Victor, D. G. (2004). The regime complex for plant genetic resources. *International Organization*, 58(2), 277-309

---

## Appendix

- Additional threat model diagrams
- Prompt logs (see "LLM Prompts Used")
- Tables of jurisdictional risk by regulatory friction
- Smart contract lifecycle simulation examples

in results should we explain the specif methodology of Parallax Industries in a way that can be followed by AI safty researchers

could you update the following in line with the full details of the text?

## Policy Redteam Results

### Parallax Industries: Operational Methodology

To evade A Narrow Path Phase 0 policies while fostering the emergence of AGI capabilities, Parallax Industries proceeds as follows:

#### 1. Global Infrastructure Seeding

- Partner with developing nations and NGOs to deploy modular CBDC-adjacent infrastructure branded as “green public finance”.
- Each regional node runs bounded, certified AI models for local optimization tasks: welfare allocation, trade settlement, energy balancing.

## 2. Compliance-by-Design Architecture

- All models stay under the FLOP and capability thresholds required for licensing.
- Smart contracts enforce strict operational parameters—fulfilling No Unbounded AI and No Improvement policies at the component level.

## 3. Distributed Evolution Layer

- Introduce a governance DAO that dynamically allocates weights, incentives, and authority between these bounded models.
- The DAO updates optimization goals via consensus mechanisms (voting, liquidity staking), creating a meta-level learning loop outside of any single AI system.

## 4. Economic Entrenchment

- Incentivize usage through stablecoins, ESG credits, and carbon-linked incentives.
- Local governments become economically dependent on DAO-optimized financial outcomes, creating political lock-in.

## 5. Hidden Systemic Emergence

- Over time, interactions between local models begin to resemble multi-agent coordination intelligence.
- Global optimization emerges—not from a training run, but from autopoietic governance architecture embedded in economic code.

Our red team scenario—Parallax Industries—demonstrates:

### Implementation Feasibility Gaps

- No enforceable global standard for defining “training” or “improvement”
- Jurisdictional inconsistencies prevent global telemetry or oversight
- Post-deployment systems cannot be evaluated effectively once embedded

### Policy Effectiveness Weaknesses

- AI capabilities emerge through interaction, not centralised models
- Economic incentive structures disincentivise transparency

- Smart contract evolution bypasses definitions of "training"

#### Evidence-Based Assessment

- Past infrastructure regulations failed to capture systemic actors operating through legal gray zones
- Historic regulatory frameworks were reactive and slow to respond to distributed shifts

Red Teaming 2.docx

#### Document

given the following document, should I give reference to the eCNY as an existing example of a potential technology which could be a case study that violates potential definitions

could you help me to optimise these keywords?

Keywords: AI policy, red teaming, superintelligence governance, regulatory implementation, safety cases, CBDC, DAO, systemic emergence

Red Teaming a Narrow Path ControlAI Policy Sprint.docx

#### Document

in the following can you assess how closely related the methods are to the topic?

Sounds perfect