Formal Models for Robust AGI Governance[1]

Geopolitics of AGI

Anton Shenk

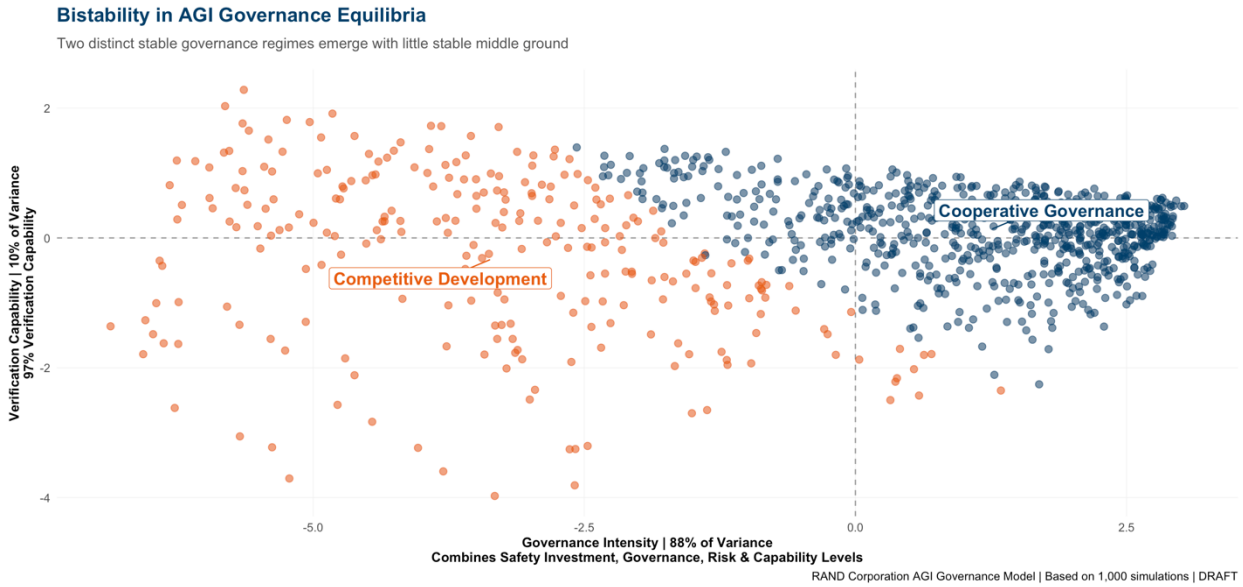
Current approaches to AGI strategy development – expert elicitation, scenario-based planning, and historical analogies – provide a valuable but incomplete analytical foundation. Where those methods excel at capturing nuance and incorporating expertise, they struggle to account for strategic interactions over time – **creating a blind spot precisely where clarity is most crucial**.

A formal game theoretic framework would provide an analytical complement by transforming implicit assumptions into explicit parameters that can be systematically varied and tested. Such a model would:

1. Expose hidden path dependencies in governance trajectories that narrative scenarios often miss
2. Isolate key parameters that disproportionately determine system behavior
3. Identify critical thresholds where incremental changes produce discontinuous outcomes

Our model demonstrates this approach in action – offering policymakers a more complete analytical toolkit for navigating among the most consequential technology governance challenges of our time.

Below, from the model, captures the essence of what modeling contributes to AGI strategy development. Each dot represents a possible stable governance outcome – clustered into two distinct states. The model identifies which combinations of parameters (verification capabilities, strength of global governance, investments in safety) produce stable outcomes – **and which create volatile conditions where small perturbations trigger systemic collapse.**



This approach doesn't just tell us what futures we might desire – it reveals which futures are achievable and how policy can help get us there.

---

[1] This unpublished, unreviewed, unedited memo was produced to share preliminary insights, hypotheses and questions emerging from the RAND Geopolitics of Artificial General Intelligence Initiative (AGI), an activity of Technology and Security Policy Center (TASP) in the RAND Global and Emerging Risks (GER) Division. The goal of the RAND Geopolitics of AGI Initiative is to develop and evaluate alternative strategies and policies for technically credible but deeply uncertain futures with AGI. Funding for this work was provided by gifts from RAND supporters. For more information about the RAND Geopolitics of AGI Initiative, please contact Jim Mitre (jmitre@rand.org) or Joel Predd (jpredd@rand.org). Please do not distribute or cite.

Approach

Our model (see *Appendix A*) formalizes AGI governance as a dynamic system of interacting nations developing capabilities while navigating safety tradeoffs. The system captures five core mechanisms:

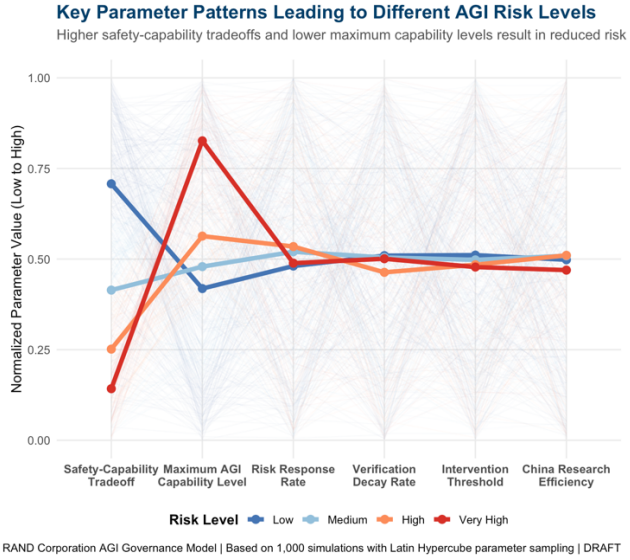
- Technological development following logistic S-curves
- Safety investments that slow capability growth
- Verification capabilities that deteriorate without maintenance
- Governance strength that responds to collective safety commitments
- Strategic interventions when capability gaps emerge

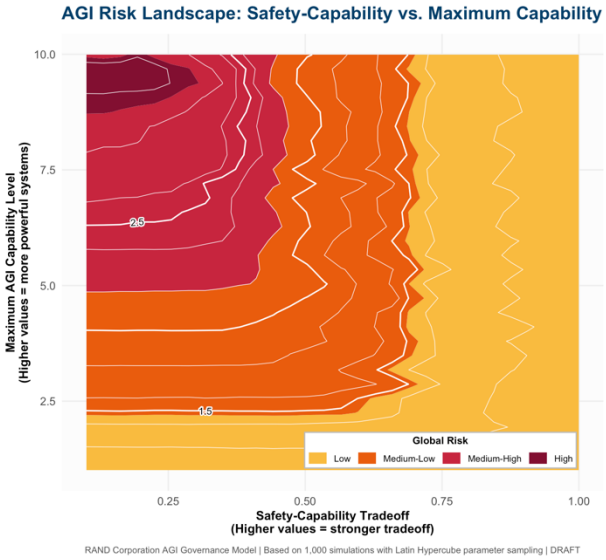
By structuring these relationships as differential equations, we transform qualitative assumptions about AGI dynamics into quantifiable parameters that can be systematically varied. This enables robust exploration of governance futures and identification of robust policy interventions that remain effective under uncertainty.

Early Results and Policy Implications

Analysis of 100,000 simulated governance years reveals counter-intuitive strategic insight – and identifies high-leverage policy interventions with potential for outsized impact.

1. Analysis reveals that the relationship between **safety research** and **development speed** is the single most influential factor affecting global risk. This suggests policymakers should prioritize governance mechanisms that maintain *meaningful* coupling between safety and capability development – rather than "safety without slowdown" futures that inadvertently enable dangerous racing dynamics.
2. Our model reveals **non-linearities in risk dynamics** – with risk not increasing gradually but rather exhibits abrupt "cliff edges" when capability gaps exceed critical thresholds. This creates narrow but consequential windows where policy can prevent systems from tipping into high-risk regimes.





Next Steps

Future work should extend this model to capture additional strategic dimensions – including multi-actor dynamics beyond the two-power framework, asymmetric information conditions, and varied technological diffusion rates. Integration with empirical data from historical technology races could help calibrate key parameters and validate behavioral assumptions. The most promising direction involves developing an interactive interface allowing policymakers to directly explore intervention effectiveness across diverse scenarios, transforming abstract insights into concrete decision support. By systematically connecting theoretical governance models with practical policy levers, this approach can help bridge the gap between policy analysis and operational strategy in this critical domain.

APPENDIX A: Simulation Parameters

Here I present a possible specification for a game theoretic model to explore AGI governance dynamics.

Capability Development

Each nation's AGI capability is a function of its research efficiency ($\mathbf{r_i}$) moderated by an S-curve, safety investments, and interventions from other nations.

$$\frac{dC_i}{dt} = \mathbf{r_i} \cdot \underbrace{\left(1 - \mathbf{s_p} \cdot \mathbf{S_i}\right)}_{Safety\ Penalty} \cdot \underbrace{\mathbf{C_i} \cdot \left(1 - \frac{\mathbf{C_i}}{C_m}\right)}_{S-Curve} - \underbrace{\sum_{j \neq i} I_{ji} \cdot \left(\mathbf{C_i} - \mathbf{min}(\mathbf{C_i}, \mathbf{C_j})\right)}_{Intervention}$$

Safety Investment

Nations adjust safety investments based on their perceptions of global risk and international governance.

$$\frac{dS_i}{dt} = \underbrace{\mathbf{r_r} \cdot (\mathbf{R} - \mathbf{S_i})}_{Risk\ Signal} + \underbrace{g_r \cdot (\mathbf{G} - \mathbf{S_i})}_{Norm\ Pressure}$$

Verification and Governance

Verification capabilities ($\mathbf{V}$) improve with governance ($\mathbf{G}$) but naturally decay over time. Governance adapts to the average safety commitments across nations, creating a collective action dynamic.

$$\frac{dV}{dt} = \underbrace{\mathbf{v_i} \cdot \mathbf{G}}_{Boost} - \underbrace{\mathbf{v_d} \cdot \mathbf{V}}_{Decay} \qquad \frac{dG}{dt} = \mathbf{g_a} \cdot \left(\frac{\sum_i S_i}{\mathbf{n}} - G\right)$$

Intervention

Nations may strategically intervene to reduce capability gaps when a competitor pulls ahead. Better verification ($\mathbf{V}$) capabilities reduce the effectiveness of such interventions.

$$I_{ij} = \mathbf{max}(\mathbf{0}, \mathbf{i_i} \cdot \left(\mathbf{C_j} - C_i - i_t\right) \cdot (\mathbf{1} - \mathbf{V}))$$

Global Risk

Global risk increases when squared capabilities outpace safety investments, creating a nonlinear relationship where advanced capabilities generate disproportionately greater risks without sufficient safety measures.

$$\frac{dR}{dt} = \mathbf{r_f} \cdot \left(\frac{\sum_i C_i^2}{\sum_i S_i \cdot \mathbf{C_i} + \varepsilon} - R\right)$$

Simulation

Latin Hypercube Sampling reveals stable regimes over 100,000 simulated governance years.

Parameter	Symbol	Range	Description
Research Efficiency (US)	r_{US}	0.1 - 0.4	Rate US converts resources into capability growth.
Research Efficiency (CCP)	r_{CCP}	0.1 - 0.4	Rate China converts resources into capability growth.
Maximum Capability	C_m	1.0 - 10.0	Theoretical ceiling for AGI capability development.
Safety Penalty	s_p	0.1 - 1.0	How much safety investments slow capability development.
Risk Response	r_r	0.05 - 0.3	How quickly nations adjust safety based on perceived risk.
Risk Factor	r_f	0.1 - 0.4	How global risk responds to capability/safety imbalance.
Governance Response	g_r	0.05 - 0.3	How strongly nations respond to governance standards.
Governance Adjustment	g_a	0.05 - 0.3	How governance adapts to safety investments.
Verification Improvement	v_i	0.1 - 0.5	Rate at which governance translates to verification capability.
Verification Decay	v_d	0.05 - 0.2	Rate at which verification capabilities deteriorate over time.
Intervention Intensity	i_i	0.1 - 0.9	Willingness to intervene against competitors' capabilities.
Intervention Threshold	i_t	0.1 - 0.5	Capability gap required for intervention.