# System Dynamics Game-Theoretic Model of the AI Development Race[1]

David Bravo
IQS

Publius Dirac
BlueQubit

Areeb Fazil
Universitat Pompeu Fabra

Ariel Gil
AI Standards Lab
Trajectory Labs

Pablo Rosado
Our World in Data

**With**
Apart Research

## Abstract

We present a game-theoretic model of AI development competition between three major actors (US, China, EU) that captures strategic tradeoffs between capability acceleration, safety investment, and international verification, with the assumption of iterative best action. The model features two-phase capability growth (diminishing returns transitioning to recursive self-improvement at AGI thresholds), safety spillovers mediated by trust, and Nash equilibrium computation via iterative best response. In some simulations, we find that knowledge of other bloc capabilities delays the decision to race (accelerate AI development), contradicting the Armstrong et al. 2016 result where knowledge of competitor capabilities worsens race dynamics. Further tuning/validation of model parameters is left for later research.

*Keywords: AI forecasting, game theory, multi-agent systems, AI safety, strategic competition, scenario analysis*

[1] Research conducted at Apart Research's AI Forecasting Hackathon, 2025. Not conducted at nor funded by the affiliated organizations.

* Authors ordered alphabetically

# 1. Introduction

We were initially interested in extending the Armstrong et al., 2016 paper, to see how the game-theoretic behavior of the different actors would change in various scenarios. We also wanted to model system dynamics of the arms race somewhat, inspired by (Wilmes & van Waas, 2024). While acknowledging that actors may not behave in a game-theoretic optimum way (Ellsberg, 2017), we still think a model might capture some important aspects of how the AGI race might play out. For instance, if AI systems are used more in decision making, we may see dynamics closer to ideal rational agents. Note that we consider multi-agent issues (Hammond et al., 2025) mostly out of scope.

## Research Questions

We investigate how game-theoretic maximizing behavior among leading AI-developing blocs shapes the trajectory toward transformative AI. Specifically, we ask:

- **Race dynamics:** How do strategic interactions between the US, China, and the EU affect timelines to transformative AI capabilities?
- **Safety incentives:** Under what conditions do actors prioritize safety investment over capability acceleration?
- **Verification and trust:** How does the strength of verification mechanisms and mutual trust influence cooperation and safety spillovers?
- **Governance regimes:** How do different governance scenarios—arms race, partial cooperation, or full coordination—affect global safety outcomes and capability timelines?
- **Sensitivity and leverage:** Which structural parameters (e.g. trust decay, safety tax, capability threshold) most strongly determine whether the system converges toward cooperative or competitive equilibria?

These questions aim to clarify when international competition amplifies existential risk and when coordination can sustainably align incentives.

## Contribution

Our main contribution is to formalize and simulate the strategic dynamics of AI development within a system-dynamics and game-theoretic model, expanding on and replicating the Armstrong et al., 2016 paper. It models the interaction between technological capability, safety effort, and institutional trust.

## Threat Model

The basic threat model is race dynamics, where overall safety is determined by the least safe actor — and so, if all actors are racing to be the "first to the finish line" (AGI/ASI), they will not invest sufficiently in safety. This is both due to resource allocation, and due to safety training sometimes being detrimental to capabilities ("safety tax"). For example, currently CoT is helpful for both capability and monitorability (Emmons et al., 2025), but in the future, it may trade off performance against more advanced architectures (Oscar, 2025).

One danger is that after a certain threshold, models may go into a self-improvement loop, where safety efforts will not keep up with the increasingly capable (and dual use) model capabilities.

## 2. Methods

At a high level, our model treats AI development as a continuous competition between three actors (the US, China, and the EU) who must decide how to allocate their limited resources between three goals: building more powerful AI systems, keeping those systems safe, and maintaining mutual trust through verification.

Each actor's choices influence not only their own progress but also the others': capability races accelerate global progress but can widen the safety gap, while investments in trust and verification can make safety research more "public" by allowing actors to benefit from each other's advances.

Over time, these feedbacks can push the system toward two different futures:

- A **race dynamic**, where distrust leads to underinvestment in safety and earlier — but riskier — AGI timelines, or
- A **cooperative dynamic**, where high trust and effective verification sustain shared safety growth and slower, safer capability advancement.

### Model Architecture

#### State Variables

Capability rises when a bloc spends effort on acceleration. Before it crosses a critical threshold, returns diminish; after that threshold, gains accelerate sharply. A smooth transition function links the two regimes so there's no sudden jump.

Safety grows through deliberate safety work and also benefits from spillovers: when global trust is high, blocs can borrow indirectly from each other's safety advances.

Trust itself accumulates when blocs invest in verification/cooperation and naturally decays if neglected. Because trust lubricates safety spillovers and yields a direct payoff bonus, sustaining it matters for everyone.

### Model Equations

For bloc $i$, capability $K_i$ follows a two-regime growth law that blends diminishing returns with recursive self-improvement:

$$\frac{dK_i}{dt} = (1 - w_i)\,\frac{\alpha\, aX_i}{1 + \beta_{\text{dim}}\, K_i} + w_i\,\alpha\, aX_i\, K_i,$$
$$w_i = \tfrac{1}{2}\left[1 + \tanh\!\left(\frac{K_i - K_{\text{threshold}}}{\text{transition\_width}}\right)\right].$$

Here $\alpha$ converts acceleration effort $aX_i$ into capability gains; $\beta_{dim}$ shapes the diminishing-returns regime before the threshold, $K_{threshold}$, which marks the onset of recursive self-improvement; and *transition_width* controls how smoothly the model moves from one regime to the other through the weighting term $w_i$.

Safety stock $S_i$ evolves through direct safety work and cooperative spillovers:

$$\frac{dS_i}{dt} = \gamma\, aS_i + \eta\, T\, \frac{1}{N-1} \sum_{j \neq i} S_j.$$

The coefficient $\gamma$ turns safety effort $aS_i$ into additional safeguards, while $\eta$ measures how effectively global trust $T$ allows bloc $i$ to benefit from the other blocs' safety stocks. The summation averages the safety levels of the other ($N$ - *1*) blocs, so higher trust makes everyone's safety work more mutually reinforcing.

Trust itself responds to verification effort and decays when neglected:

$$\frac{dT}{dt} = \beta\, \frac{1}{N} \sum_{i=1}^{N} aV_i - \delta_T\, T.$$

The parameter $\beta$ captures how rapidly shared verification/cooperation effort $aV_i$ builds trust, whereas $\delta_T$ is the natural decay rate if no one continues investing. Sustaining trust keeps the spillover channel open and, in the payoff model, yields a direct cooperative dividend.

Safety debt of bloc $i$ measures how much capability outpaces safeguards:

$$\text{Safety\_debt}_i = \max\left(0,\ K_i - \theta\, S_i\right)$$

The parameter $\theta$ tells how effectively safety mitigates capability risk. When safety keeps pace, debt is zero; otherwise debt grows in proportion to the uncovered capability. We track this debt to flag when any bloc's technological lead becomes dangerously undersecured.

Each bloc's instantaneous payoff combines capability benefits, penalties for debt, and the cooperative upside of trust:

$$\text{Payoff}_i = K_i - \lambda\,\text{Safety\_debt}_i + \omega\,T$$

The first term captures the direct value of capability. The second term subtracts a fraction $\lambda$ of safety debt, reflecting how risky overhang erodes utility (higher $\lambda$ means stronger aversion to debt). The final term adds a trust dividend scaled by $\omega$; global trust ($T$) lowers coordination frictions and keeps spillovers active, so we treat it as a shared public good that boosts everyone's payoff.

As well as the instantaneous payoff, a look-ahead parameter is added which accounts for the expected reward of each block for the following one year. In the "Bostrom Minimal" scenario (detailed below), this parameter is not used.

#### Simulation Workflow

Parameter values and initial conditions are supplied either from calibration data or from intuitive estimates. The model then integrates the system of differential equations forward in time to trace capability, safety, and trust trajectories.

During the run, the policy logic is reapplied at each step to record the implied actions, enabling diagnostics like the evolution of safety debt.

#### Parameters

The simulation parameters were estimated based on real-world data from Epoch AI, with model compute used as a proxy for per-bloc starting capability. Other parameters were set approximately, without much calibration to real data.

We focused on getting the model to show somewhat interesting dynamics with starting parameters, and leave as next steps to ground them further.

### Experimental Design

For simplicity, we modeled two discrete scenarios:

1. **Bostrom Minimal:** Each bloc maximizes only its current payoff, blind to rivals' incentives and future impacts; this is the short-sighted race baseline.
2. **Public Information:** Blocs can view everyone's payoffs and simulate outcomes one year ahead, so they respond to anticipated shifts in capability, trust, and debt.

## 3. Results

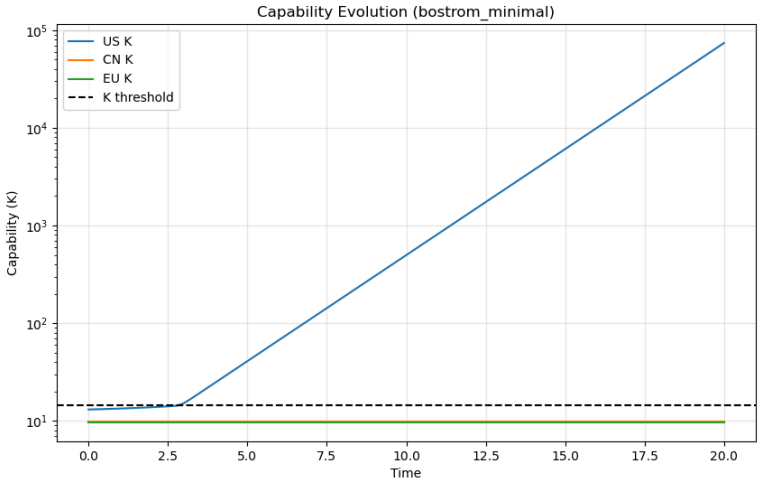
The Armstrong et al. (2016) paper finds that knowledge of competitor capabilities worsens race dynamics. We test this in our simulation, and under some parameters

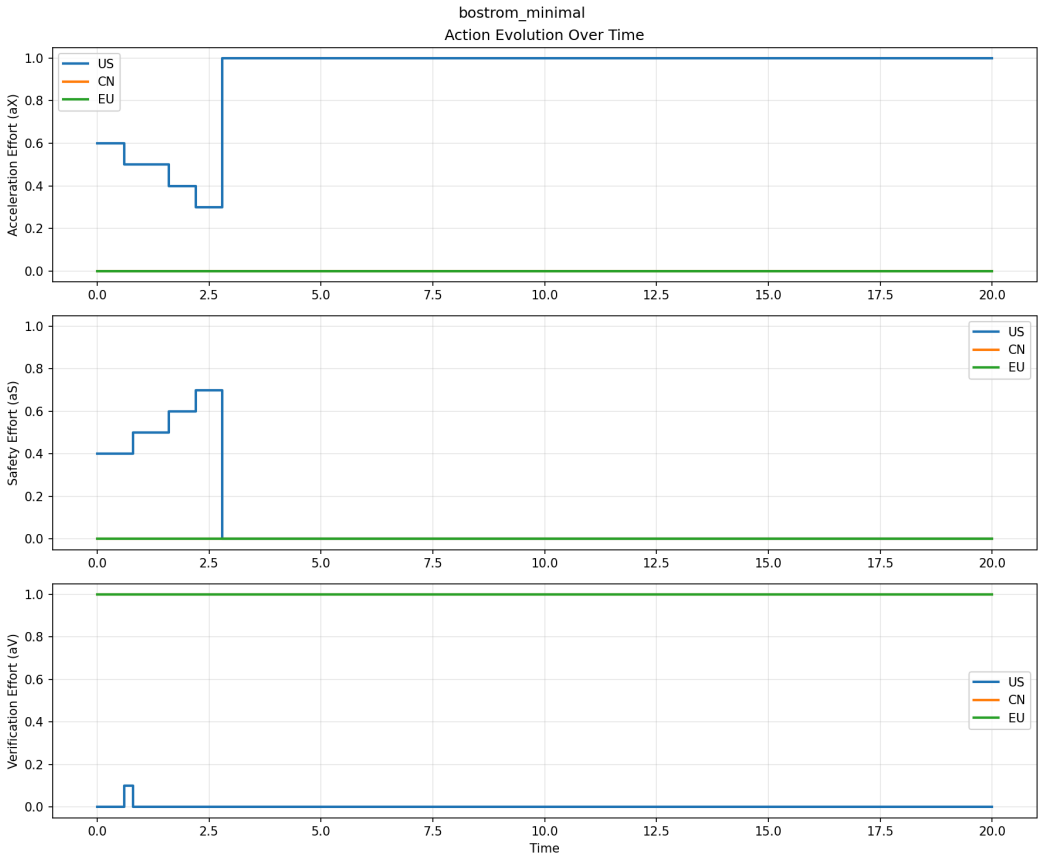
we find that knowledge of other bloc capabilities delays the US decision to race (accelerate AI development). In the Bostrom Minimal scenario, the US switches to "accelerate" at year 2.5, while the Public Information scenario shows the US accelerating at year 11 of the simulation.

With different starting parameters, we did not see this behavior; suggesting it may be sensitive to parameters. Still, this is an interesting result and we think it is worth exploring further as the parameters are further tuned.

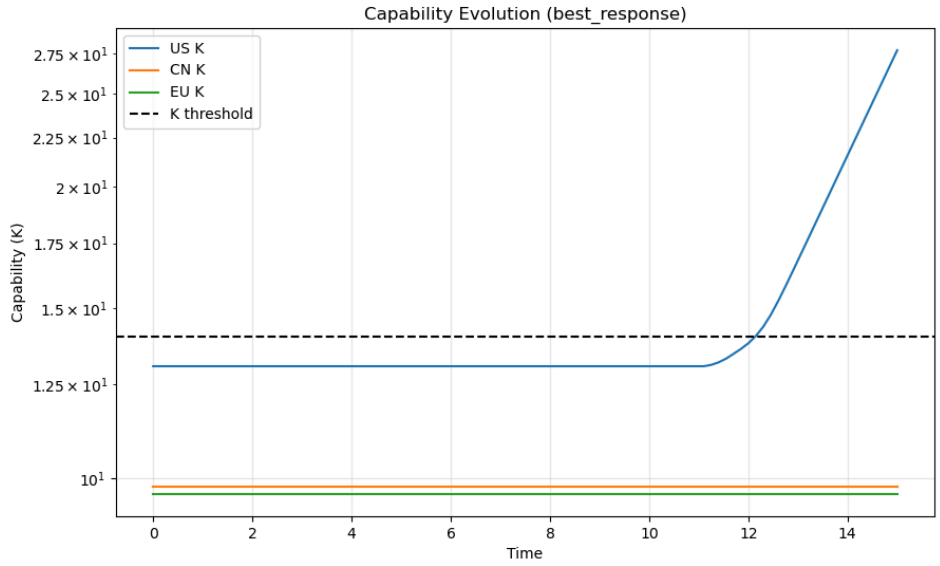
Plots of this behavior are shown below.

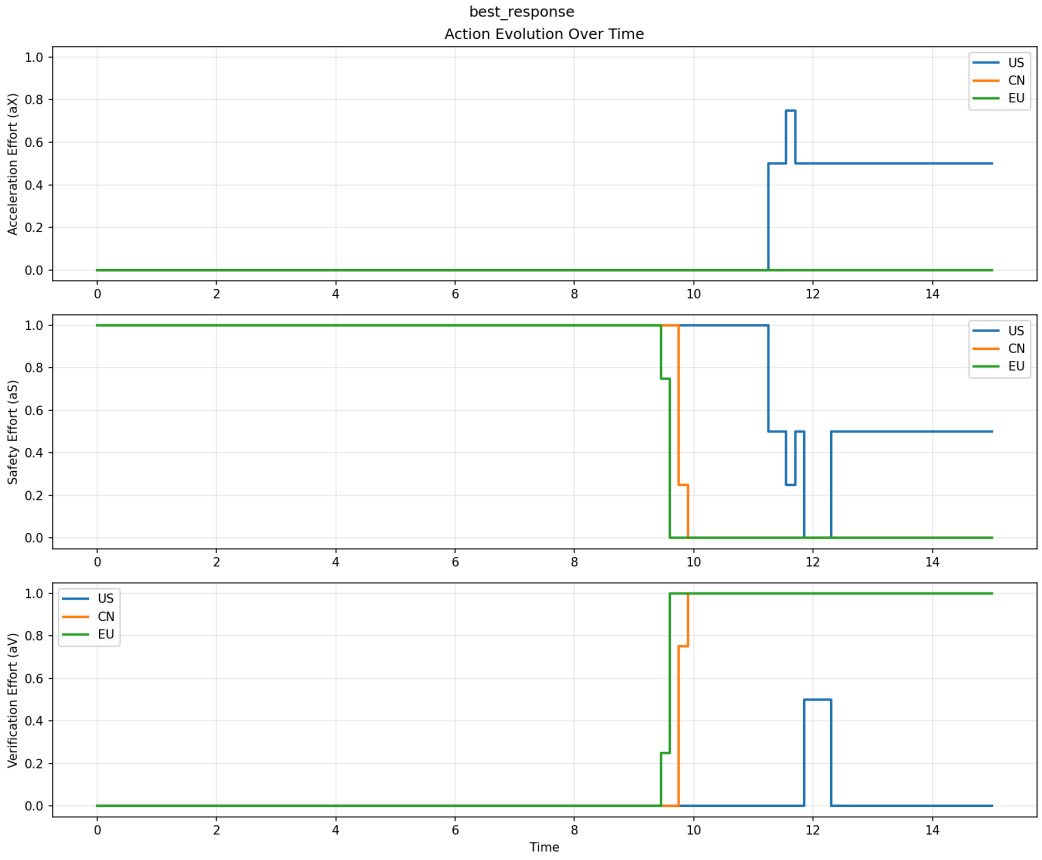
## Scenario 1: Bostrom Minimal



## Scenario 2: Public Information

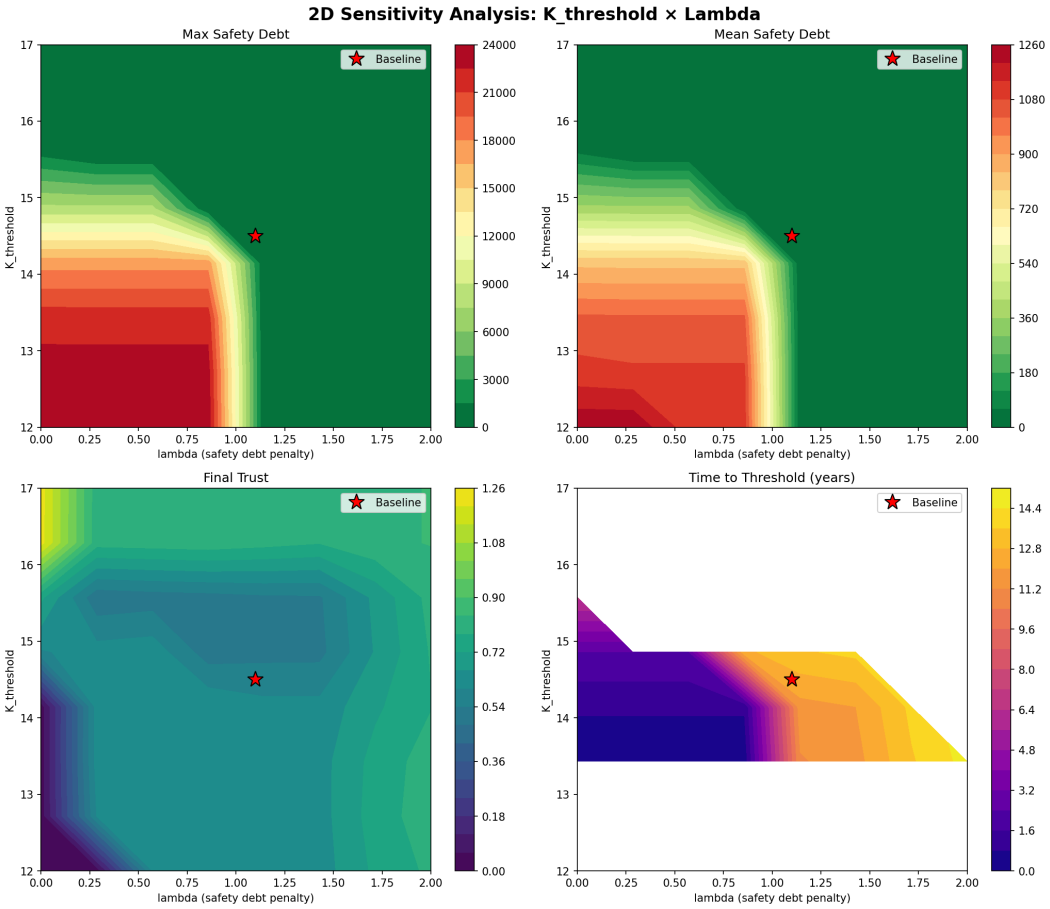



## Sensitivity Analysis

We performed an initial sensitivity analysis for parameter $K_{threshold}$, which represents the critical threshold of capabilities above which capability gains

accelerate sharply; and $\lambda$, which reflects the penalty of the safety debt on the payoff.

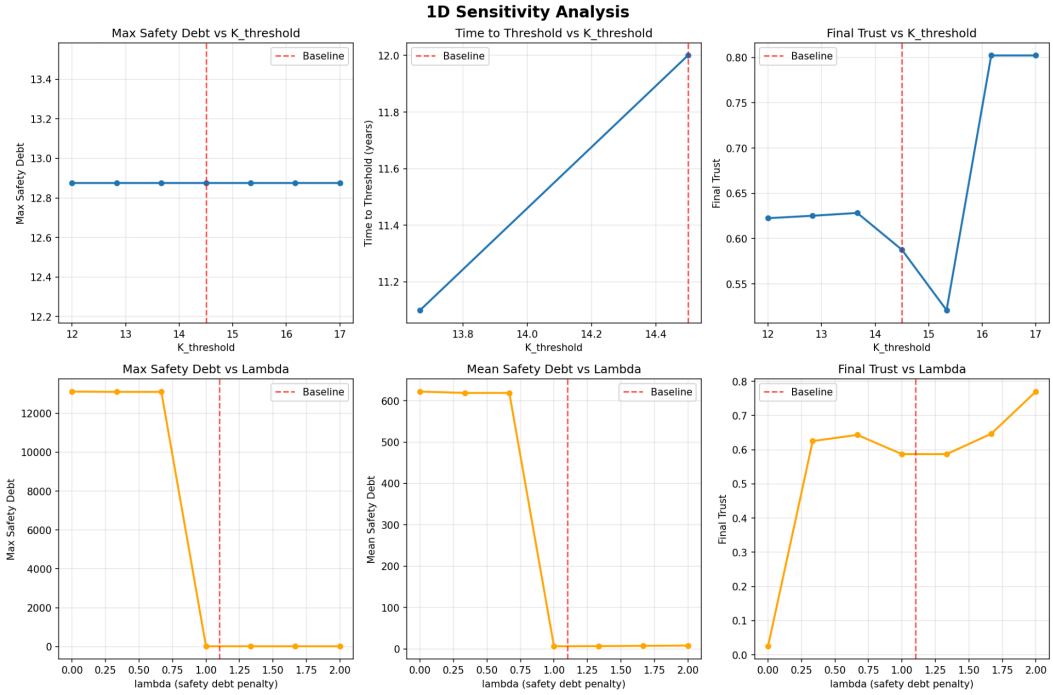



Top left: Max Safety Debt (color) plotted against K_threshold (capability at which point self improvement starts) and lambda (Safety Debt Penalty to the payoff)

Top Right: Mean Safety Debt (color) plotted against K_threshold and lambda,

Bottom Left: Final Trust - trust level at the end of the simulation (color) plotted against K_threshold and lambda.

Bottom Right: Time to Threshold (color) - simulation year at which K_threshold is reached




## 4. Discussion and Conclusion

This work presented a game-theoretic model of AI development competition between three major actors (US, China, EU) that captures strategic tradeoffs between capability acceleration, safety investment, and international verification.

While we saw some interesting differences between access to other actor capability information, we believe the parameters of the model are still preliminary. Still, we think that using system dynamics is promising - it allows flexibility in modeling different actors and their actions iteratively, and lets us simulate complex payoffs for each actor very quickly - 1-2 minutes for a full simulation.

We believe this model can serve as inspiration for more informative AI timeline and risk modeling.

## 5. Acknowledgements

This work was assisted by Allen Abishek, EA Barcelona, and AI Safety Barcelona.

## 6. References

Armstrong, S., Bostrom, N., & Shulman, C. (2016). Racing to the precipice: A model of artificial intelligence development. *AI*Armstrong, S., Bostrom, N.,

& Shulman, C. (2016). Racing to the precipice: A model of artificial intelligence development. *AI & SOCIETY*, *31*(2), 201–206. https://doi.org/10.1007/s00146-015-0590-y

Ellsberg, D. (2017). *The doomsday machine: Confessions of a nuclear war planner*. Bloomsbury.

Emmons, S., Jenner, E., Elson, D. K., Saurous, R. A., Rajamanoharan, S., Chen, H., Shafkat, I., & Shah, R. (2025). *When Chain of Thought is Necessary, Language Models Struggle to Evade Monitors* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2507.05246

Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., Smith, C., Barfuss, W., Foerster, J., Gavenčiak, T., Han, T. A., Hughes, E., Kovařík, V., Kulveit, J., Leibo, J. Z., Oesterheld, C., de Witt, C. S., Shah, N., Wellman, M., ... Rahwan, I. (2025). *Multi-Agent Risks from Advanced AI* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2502.14143

Oscar. (2025, September 26). On keeping chains of thought monitorable. *LessWrong*. https://www.lesswrong.com/posts/3Kpn3Ea4x6tgdnW7R/on-keeping-chains-of-thought-monitorable

Wilmes, L., & van Waas, R. (2024). Understanding Arms Races for Autonomous Military Capabilities Using a System Dynamics Simulation Model. *NATO Journal of Science and Technology. & SOCIETY*, *31*(2), 201–206. https://doi.org/10.1007/s00146-015-0590-y

# 7. Appendices

## A. Security Risks

This model is a rough approximation of race dynamics. The parameters are not yet very accurate, and so in the wrong hands, the conclusions drawn from the model could lead to incorrect policy actions.

More detailed tuning of the model parameters is left for future work - we observed very different behavior when varying the parameters.

The model is also only simulating game theoretical agents, rather than real world dynamics. The game theoretic optimal actions are not necessarily the most likely real world actions.

## B. Limitations

Even as the model helps illuminate race/cooperation feedback, it rests on multiple simplified assumptions that bound how much we can trust its projections. The main caveats are summarized below.

- Dynamics are fully deterministic, so the model cannot capture shocks, crises, or the stochastic breakthroughs that often shape real AI trajectories.
- Every bloc shares the same functional form and action budget, even though actual states have asymmetrical capabilities, resources, and institutional constraints, and the generic budget is not tied to any real economic capacity.
- Payoffs compress complex incentives into a single formula that values capability, penalizes "safety debt," and rewards shared trust. In practice, payoffs depend on domestic politics, regulatory pressures, and risk tolerance that vary wildly across actors.
- The steep capability threshold is also speculative: the *tanh* transition is meant to mimic recursive self-improvement, but its location and width are assumptions rather than empirically grounded facts, so the model is sensitive to unverified tipping points.
- Trust and spillovers are modeled with simple averages, and the iterated best-response routine is an abstraction of strategic behavior. Real negotiations, coalitions, and verification regimes are lumpy, path-dependent, and rife with misperception.
- Parameter values throughout the system come from judgement calls rather than statistical calibration, so any quantitative forecast should be treated as illustrative

## C. Future Improvements

Future work should calibrate the model to empirical series such as the Stanford AI Index and Epoch AI compute estimates so capability growth, safety investments, and trust decay reflect observed trends. With firmer grounding, users could stress-test specific policy interventions (compute caps, mandated safety audits, or alliance-based verification regimes) and quantify their impact on capability balances and safety debt.

Adding stochastic shocks and interactive policy controls would push the framework closer to decision support. Unexpected breakthroughs, accidents, or trust crises could reveal how fragile equilibria are, while a live interface would let analysts adjust interventions on the fly and explore counterfactual timelines with mixed deterministic and probabilistic elements.