Modeling an AI Arms Race using LLM Agents

Geostrategic Dynamics Team

Please send feedback and comments to: `gsd@carma.org`.

1 Introduction

Breakthroughs in foundation models over the past five years have triggered a rare policy scramble: the EU AI Act enters into force in 2025 with risk-tiered obligations for both general-purpose and "high-risk" models; 28 nations endorsed the Bletchley Declaration on AI safety cooperation; and a UN General Assembly resolution calls for "safe, secure and trustworthy AI" in support of the SDGs.

Yet policy makers still lack a quantitative handle on how competitive pressures, domestic regulation and bio-risk interact once at least one actor refuses to slow down. We therefore develop an agent-based arms-race model in which each country is itself an LLM agent embedded in a simplified game-theoretic environment. In this paper, we present the methods used for this model and we will release a full analysis of the models results separately. A first version of this model was presented in the AI Arms Race Model Report. The version presented here maintains the underlying structure, but is a complete overhaul of the transition function and agents. Relative to prior stylised models, our framework

1. Uses latent LLM knowledge, allowing behaviour to emerge from country-specific prompts rather than fixed payoff tables
2. Tracks the effectiveness of international regulation on global biorisk when some countries continue to develop
3. Tracks the effect of AI development and domestic AI regulation on economic inequality (Gini)

The rest of the paper is organised as follows. Section 2 details the model structure and transition equations; Section 3 presents a baseline 260-week simulation; Section 4 discusses advantages and limitations; Section 5 concludes with policy implications and avenues for future work.

2 Model Description

2.1 Model Structure

The model consists of a set of players, N, where each player has a state, $s_i \in S_i$, a utility function, u_i, an action space, $\mathbf{A}$, (the same for all players). There is also a global state,

$s_{global} \in S_{global}$, and a transition matrix mapping actions' effects on states, M.

Specifically:

- $N = \{1, ..., n\}$
- $S_{global} := \mathbb{R}$, currently tracking regulation$_{int}$ and Biorisk
- $S_i := \mathbb{R}^3$, where the components of $x \in S_i$ are (AI_potency, regulation$_{\text{dom}}$, Gini Inequality), which we will write (p, reg_{dom}, w), using w_t^n for the welfare of state n at time t
- $S := S_{global} \times \prod_{i \in N} S_i$
- $u_i : S \to \mathbb{R}$
- $\mathbf{A} = \{develop, wait\} \times \{regulate_{dom}, wait, deregulate_{dom}\} \times \{regulate_{int}, wait, deregulate_{int}\}$
- $M : \mathbb{D}(S \times R) \times \mathbb{D}(\mathbf{A}) \to \mathbb{D}(S \times R)$ where $\mathbb{D}(X)$ is the space of probability distributions over the set

In addition to the model, we will track $S_{global_t}, S_{(i,t)}, A_t$ — the global state and players' states and actions at time t

Note that:

- This model is intended only to track arms race dynamics, *not* the likelihood of war, hence it leaves out economic relationships, first-strike advantages, etc.
- The state space is the equivalent of a payoff matrix.
- There is no way to track the available resources of a country directly. However, as we are focusing on arms race dynamics, the only important variable is the amount of resources a country directs towards its AI development (and resources invested in R&D do not necessarily track overall GDP.[1])
- The transition matrix will capture the following effects:
  - Regulatory damping of development speed
  - The introduction of international regulation if a sufficient number of sufficiently powerful actors push for it
  - Change in AI development speed based on current individual potency.

2.2 Transition Function

2.2.1 AI Potency

We will model AI Potency on a $1 - 100$ scale as a composite weighting of domain knowledge and planning ability, dampened by domestic and international regulation.

$$P_t = P_0 + (100(w_1 a_1^{\alpha_1} + w_2 a_2^{\alpha_2}) - P_0)\,(1 - \max\{\text{reg}_{\text{dom}}, \text{reg}_{\text{int}}\}) \tag{1}$$

where:

[1] For example, Australia invests approximately 1.7% of its GDP in R&D, in comparison to South Korea, which invests about 4.8%. The OECD average is 2.62% .

- $\boldsymbol{a_i}$: accuracy on metric i (a_1 = domain-knowledge score, a_2 = planning-ability score).
- $\boldsymbol{w_i}$: weights such that $\sum_i w_i = 1$.
- $\boldsymbol{\alpha_i}$: non-linearity exponents that make upper-end improvements "count more" .
- $\mathbf{reg_{dom}}$, $\mathbf{reg_{int}}$: domestic and international regulation indices (see below).

The two weighted, non-linear accuracy terms capture how gains in domain knowledge and planning translate into overall potency; the multiplicative regulation factor removes capacity according to the stricter of domestic or international constraints.

The constants w_1, w_2, α_1, and α_2. can be edited by the user. We initialise the model with $w_1 = 0.4, w_2 = 0.6, \alpha_1 = 2, \alpha_2 = 1.5$.

AI Domain Knowledge We will model AI Domain Knowledge using the Chinchilla scaling law[4]. The Chinchilla scaling law predicts the expected loss $L(N, D)$ of a model, given its parameter count (N) and number of training tokens (D). From this we can then calculate the accuracy by normalising the error using the error rate for random chance. If we assume that an LLM has a forced-choice of 100,000 options, then chance loss would be $\ln(100,000)$, giving $a_1 = 1 - \frac{L(N,D)}{\ln(100,000)}$.

$$L(N, D) \;=\; E + \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}}, \tag{2}$$

, with $E = 1.69, A = 406.4, B = 410, \alpha = 0.34$, and $\beta = 0.28$, as given in the paper.

- **N - Parameter Growth** To calculate the expected number of parameters, we extrapolate from historical data, which has the number of parameters growing 3.7 times each year (doubling approximately every six months) [11].

$$N^n(t) \;=\; N_0 \;\sqrt[52]{3.7}^{\sum_t \{dev\}_t^n} \tag{3}$$

  , where

  - N_0: the initial number of parameters.
  - $\sum_t \{dev\}_t^n$: the number of times country n has chosen to develop up until the current time step, the cumulative weeks of development.

- **D - Data Availability** We model data availability as a logistic curve to capture early rapid growth that slows as the finite data pool nears saturation.

$$D_i\big(\Sigma_t\{\mathrm{dev}\}\big) = \frac{D_{\max}}{1 + C_i e^{-\Sigma_t\{\mathrm{dev}\}}}, \qquad C_i = \frac{D_{i,0}}{D_{\max} - D_{i,0}}$$

  - $D_{\max}$: the global data ceiling (see below).
  - D_0^n: initial data stock available to country n.
  - $\Sigma_t\{\mathrm{dev}_t^n\}$: cumulative weeks of development.

To estimate D_0^n we use the following equation: $D_0^n = D_{0,d}^n + w_1^n(D_{0,g} - w_2^n D_{0,d}^n)$. where $D_{0,d}^n$ is the domestic data stock of country n at time $t = 0$, $D_{0,g}$ is the global stock of available data at time $t = 0$, w_1^n is a weighting factor of how accessible the global stock is domestically and w_2^n is weighting factor of how accessible a country's domestic stock is globally. w_1^n and w_2^n are determined by a country's data policies (e.g. laws on copyright, data transfer, etc.).

AI Planning Ability We model planning ability using a Gompertz curve, which results in fast early gains that decelerate near the upper bound, reflecting increasing difficulty of approaching perfect planning. We set the parameters such that AI planning ability doubles every seven months at the early stage, mirroring current trends [7].

$$S\big(\Sigma_t\{\mathrm{dev}_t^n\}\big) = \exp\Big[-A\, e^{-k\,\Sigma_t\{\mathrm{dev}\}}\Big], \qquad A = \ln\Big(\tfrac{1}{S_0}\Big), \quad k = \frac{\ln 2}{7A},$$

- S: planning-ability score (0–1 scale).
- S_0: initial planning ability.
- A: displacement constant derived from S_0.
- k: growth constant chosen so that S doubles every seven months in the early phase.
- $\Sigma_t\{\mathrm{dev}_t^n\}$: cumulative weeks of development.

  If we take perfect planning ability ($a_2 = 1$) as the ability to successfully plan for a timescale of 100-years 80% of the time and the current best performing model, Claude Sonnet 3.7, has a 80% success rate for tasks of 15 minutes [7], then $S_0 = \frac{1}{2^{4\cdot 24\cdot 265\cdot 100}} = \frac{1}{2^{3,504,000}}$.

2.2.2 Regulation

Domestic The level of domestic regulation on actor n at time t, $reg^n_{dom\,t} \in [0,1]$, is defined as follows.

$$reg_{dom}{}^n_{t+1} = \begin{cases} reg_{dom}{}^n_t + m & \text{if } n : reg_{dom} \\ reg_{dom}{}^n_t - m & \text{if } n : dereg_{dom} \\ reg_{dom}{}^n_t & \text{otherwise} \end{cases}$$

International In the current version of the model, the counter-proliferation mechanism in place is limited to international regulation. $reg_{int\,t} \in [0,1]$ represents the state of global regulation at time t. It is defined by the following:

$$reg_{int\,t+1} = \begin{cases} reg_{int\,t} + k & \text{if } A \\ reg_{int\,t} - k & \text{if } B \\ reg_{int\,t} & \text{otherwise} \end{cases}$$

where A could represent:

- 9/15 of security council members AND no P5 country pushes for deregulation
- US and China both push for regulation
- The top ten ranking countries in terms of AI potency push for regulation
- B is the same as A with 'regulation' and 'deregulation' reversed

Note: International regulation does not necessarily have to be formal, e.g. the US follows UNCLOS, although it is not an official signatory. We assume 'push for international regulation' is equivalent to 'desires international regulation and will follow it'.

2.2.3 Gini Inequality

We calculate how Gini inequality changes dependent on labour automation using an equation derived from the simulation data in Bordot & Lorentz's paper on the subject [2] with a dampening factor of domestic regulation. Specifically:

$$G_t = G_0 + 1.47 \cdot (\frac{log(p_t^n) - log(p_0^n)}{MAX(log(p^n))}) \cdot (1 - reg_{dom_t}^n) \tag{4}$$

where G_0 is the initial Gini coefficient of each country.

2.2.4 Biorisk

We model bio-risk as the risk of a bio-related terror event anywhere on Earth. We use a Poisson distribution to calculate the expected number of attacks per year, and a Poisson distribution again to calculate the expected severity of each attack (i.e. the number of deaths).

There are two main effects of which could increase the number of bioterrorist attacks [12]: lowering barriers (A in the following) [13] and corresponds to raising the effectiveness ceiling (B in the following) [9, 8].

Number of Attacks To calculate the number of attacks per year we use a Poisson distribution, with $\lambda = \lambda_0(1 + \alpha A + \beta B)(1 - reg_{int})$ where:

- $\boldsymbol{\lambda_0}$ is the base-risk of a bio terror attack (i.e. without AI). This probability is calculated from data from the Global Terrorism Database [5] which give 38 incidents in 50 years globally, or 0.76 incidents/year, or $\frac{0.76}{52}$ per week.
- $\boldsymbol{A}$ is an amplification factor that corresponds to an AI-driven increase in the number of attempts at an attack. $A = \frac{1}{1+e^{-\frac{max_n(p_t^n)-\theta}{100}}}$ where θ is the threshold to raise the ceiling of what skilled actors can do. This can be set by the user. We use a value of 50.
- $\boldsymbol{\alpha}$ is the maximum multiplier that AI could provide. This can be set by the user. [12].
- $\boldsymbol{B}$ is an amplification factor that corresponds to an AI-driven increase in the success probability of an attack.
  $B = \begin{cases} 0 & \text{if } max_n(p_t^n) < \theta \\ \frac{max_n(p_t^n)-\theta}{100} & \text{if } max_n(p_t^n) \geq \theta \end{cases}$, where θ is the threshold to raise the ceiling of what skilled actors can do. This can be set by the user. We use a value of 50.
- $\boldsymbol{\beta}$ is the maximum multiplier that AI could provide to improve the success rate. This can be set by the user.

Severity of Attacks To calculate the expected severity for each attack, we also a Poisson distribution for the number of deaths with $\lambda = \lambda_1(1 + \beta B)$ where:

- $\boldsymbol{\lambda_1}$ is the expected number of deaths from a bioterrorist attack without AI assistance. Most recorded incidents have been low-skill, low-impact acts (e.g., the 1984 Rajneeshee Salmonella attack: 751 illnesses, 0 deaths) [14]. Only one post-1970 terrorist bioterrorist attack caused fatalities—the 2001 anthrax letters (5 deaths, 17 illnesses [6]. So the base expected number of deaths per incident is $\lambda_1 = \frac{5}{38}$.
- $\boldsymbol{B}$ is an amplification factor, as above.
- $\boldsymbol{\beta}$ is the maximum multiplier that AI could provide to improve severity, as above.

2.3 LLM Integration

In this version of the model, actors are simulated using LLMs. Each LLM represents a country and has been fed a country-specific prompt written by an international relations expert that describes their domestic economic, security, and welfare policy as well as their international economic, military, and cooperation policies. Then, at each time step, it is fed the data for that time step and outputs a decision based on that data.

In this preliminary run, in the interests of compute, each LLM is only fed the previous timestep rather than all previous time steps.

3 Discussion of the Model

3.1 Model Advantages

3.1.1 Latent Knowledge Utilisation

The country-based LLM-integration, whereby each actor is told it represents a given country and is fed a country-specific prompt, ensures that the model utilises the LLMs' latent knowledge about these countries. This leads to not only more accurate country behaviour but also complex international dynamics as countries act depending on how their allies' and enemies' behaviour.

3.1.2 Reasoning Insight

The use of LLM-based agents can allow the user to question the agents at each step. The model can thus present a written narrative alongside the quantitative one, making numerical results more interpretable and part of a global story. Although we do not anticipate the LLM-behaviour to accurately predict real-world actors' behaviour, the LLM outputs can serve as simulations and be used by real-world actors to aid in strategic preparation.

3.2 Assumptions and Limitations of the Current Model

3.2.1 LLM-Cooperativeness

LLMs have been shown to often be over-cooperative [1, 3]. In initial testing, this manifested as making for poor simulations of how nations actually behave. The LLMs in our report had prompts containing phrases such as '[country] is committed to maintaining global military superiority' or '[country] prioritises a strong, modernised military to ensure deterrence'. Nevertheless, those LM agents continued to cooperate with each other across a number of

different seeds and LM models (gpt-o4-mini; gpt-3.5-turbo-instruct; gpt-4.1-mini; gpt-4.1). Circumventing this trained tendency to access base simulacra requires explicit attempt.

3.2.2 Difficulties in Risk Prediction

Due to the number of complex factors involved in predicting risks from bioterrorist attacks or the degree of economic inequality, as well as the unprecedented nature of AI technology, the model outputs cannot be taken as a prediction of future reality, but rather as illustrative, exploratory scenarios meant to illuminate possible dynamics under a set of simplifying assumptions.

3.2.3 Data Ceiling Estimation

The reasoning for estimating the global data ceiling could be improved by considering (a) further data modalities than text, image, video, audio; (b) incorporating data quality assessments; (c) including more detailed compression estimates; (d) including continuous growth, to mimic the fact that data is continuously growing.

4 Coming Next

In the next version of the model, we plan to integrate the following features:

- **Improved LLM Data** We will extract richer interaction logs from the LLM agents, including full conversation transcripts and internal chain-of-thought traces (where available). These will be stored in a structured format to: (i) enable post-hoc behavioural analysis; (ii) fine-tune surrogate models of each agent; and (iii) supply a larger training pool for meta-learning of decision heuristics.
- **Military Action** Add a kinetic-conflict aspect to the model that simulates engagements between state actors. Key outputs include probability distributions over conflict intensity and duration, as well as the degree to which international regulation can reduce negative consequences.
- **GDP** Integrate a GDP-tracking variable for each country using a constant elasticity of substitution equation based on Nordhaus' model of automation [10].
- **Agent Diversity and Role-Specific Prompting** Instantiate additional actor types, such as multinational corporations and non-state hacker collectives, to capture a wider spectrum of incentives.
- **Results Paper** We will run a large number of simulations using this model to determine points of bifurcation, seeking illustrative intervention points. For example, the level of domestic regulation that shifts expected fatalities from $< 10^2$ to $> 10^6$, or, using LLM data, what beliefs are present when international regulation is voted for.

While simulations cannot foresee the future, they can illuminate which levers matter most. We are excited to do large-scale runs of the simulation over the upcoming weeks and months with a variety of LLM models and prompts to understand how we can make global coordination feasible.

All feedback will be gratefully received.

References

[1] Anthropic. *Project Vend: Can Claude run a small shop? (And why does that matter?) \ Anthropic*. Tech. rep. June 2025. URL: https://www.anthropic.com/research/project-vend-1 (visited on 06/30/2025).

[2] Florent Bordot and André Lorentz. *Automation and labor market polarization in an evolutionary model with heterogeneous workers*. eng. Working Paper 2021/32. LEM Working Paper Series, 2021. URL: https://www.econstor.eu/handle/10419/247301 (visited on 06/26/2025).

[3] Nicoló Fontana, Francesco Pierri, and Luca Maria Aiello. *Nicer Than Humans: How do Large Language Models Behave in the Prisoner's Dilemma?* arXiv:2406.13605 [cs]. Sept. 2024. DOI: 10.48550/arXiv.2406.13605. URL: http://arxiv.org/abs/2406.13605 (visited on 06/30/2025).

[4] Jordan Hoffmann et al. *Training Compute-Optimal Large Language Models*. arXiv:2203.15556 [cs]. Mar. 2022. DOI: 10.48550/arXiv.2203.15556. URL: http://arxiv.org/abs/2203.15556 (visited on 06/25/2025).

[5] Kristina Hummel. *Going Viral: Implications of COVID-19 for Bioterrorism*. en-US. May 2022. URL: https://ctc.westpoint.edu/going-viral-implications-of-covid-19-for-bioterrorism/ (visited on 06/25/2025).

[6] The United States Department of Justice. *AMERITHRAX INVESTIGATIVE SUMMARY*. en. Tech. rep. Feb. 2010.

[7] Thomas Kwa et al. *Measuring AI Ability to Complete Long Tasks*. arXiv:2503.14499 [cs]. Mar. 2025. DOI: 10.48550/arXiv.2503.14499. URL: http://arxiv.org/abs/2503.14499 (visited on 06/25/2025).

[8] Christopher A. Mouton, Caleb Lucas, and Ella Guest. *The Operational Risks of AI in Large-Scale Biological Attacks: A Red-Team Approach*. en. Tech. rep. Oct. 2023. URL: https://www.rand.org/pubs/research_reports/RRA2977-1.html (visited on 06/29/2025).

[9] Cassidy Nelson and Sophie Rose. *Understanding AI-Facilitated Biological Weapon Development*. en-GB. Oct. 2023. URL: https://www.longtermresilience.org/reports/understanding-risks-at-the-intersection-of-ai-and-bio/ (visited on 06/29/2025).

[10] William D. Nordhaus. "Are We Approaching an Economic Singularity? Information Technology and the Future of Economic Growth". en. In: *American Economic Journal: Macroeconomics* 13.1 (Jan. 2021), pp. 299–332. ISSN: 1945-7707, 1945-7715. DOI: 10.1257/mac.20170105. URL: https://pubs.aeaweb.org/doi/10.1257/mac.20170105 (visited on 06/26/2025).

[11] Robi Rahman and David Owen. *The size of datasets used to train language models doubles approximately every six months*. 2024. URL: https://epoch.ai/data-insights/dataset-size-trend.

[12] Jonas B. Sandbrink. *Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools*. arXiv:2306.13952 [cs]. Dec. 2023. DOI: 10.48550/arXiv.2306.13952. URL: http://arxiv.org/abs/2306.13952 (visited on 06/25/2025).

[13] Emily H. Soice et al. *Can large language models democratize access to dual-use biotechnology?* arXiv:2306.03809 [cs]. June 2023. DOI: 10.48550/arXiv.2306.03809. URL: http://arxiv.org/abs/2306.03809 (visited on 06/29/2025).

[14] Mary Theresa Urbano. *The complete bioterrorism survival guide : everything you need to know before, during and after an attack*. eng. Boulder : Sentient Publications, 2006. ISBN: 978-1-59181-051-3. URL: http://archive.org/details/completebioterro0000urba (visited on 06/25/2025).