AI Arms Race Model Report

Abra Ganz[1], Daniele Palombi[1,2], and Richard Mallah[1]

[1]Center for AI Risk Management and Alignment
[2]Institute for Categorical Cybernetics

Corresponding Author: abra@carma.org.

1 Introduction

In an era defined by rapid advancements in artificial intelligence (AI), the dynamics of a potential AI arms race and where it leads should be a pressing concern for policymakers. The mitigation of such an arms race is a significant challenge, though theoretically tractable. This is not least because there are a large number of complex dynamics at play, between regulatory policies, AI development trajectories, national interests, and the broader landscape of international relations. A better understanding of these causal relationships and emergent properties can help us identify the salient levers for mitigating or preventing the harmful effects of escalating AI competition.

This report introduces an agent-based game theoretic simulation of an AI arms race. The report outlines the structure of the model, presents a preliminary version, and provides a roadmap for future development. By modelling the AI arms race in this way, we can perform sensitivity analysis to discover which inputs result in which outputs, and thus help inform where humanity's effort should be focused. The model will soon be published as a Python notebook in this GitHub repo in late 2024.

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

Specifically:

- $N = \{1, ..., n\}$
- $S_{global} := \mathbb{R}$, currently tracking $regulation_{int}$
- $S_i := \mathbb{R}^3$, where the components of $x \in S_i$ are (AI_potency, $\text{regulation}_{\text{dom}}$, welfare), which we will write (p, reg_{dom}, w), using w_t^n for the welfare of state n at time t

- $S := S_{global} \times \prod_{i \in N} S_i$
- $u_i : S \to \mathbb{R}$
- $\mathbf{A} = \{develop, wait\} \times \{regulate_{dom}, wait, deregulate_{dom}\} \times \{regulate_{int}, wait, deregulate_{int}\}$
- $R \in Mat_{|N| \times |N|}(\{Enemy, Ally\})$
- $M : \mathbb{D}(S \times R) \times \mathbb{D}(\mathbf{A}) \to \mathbb{D}(S \times R)$ where $\mathbb{D}(X)$ is the space of probability distributions over the set

In addition to the model, we will track $S_{global_t}, S_{(i,t)}, R_t, A_t$ — the global state and players' states, relationships, 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.
- Companies, as well as countries, can be included in this model. The $regulation_{dom}$ variable for a company will simply track $regulation_{dom}$ for the country in which it is domiciled.
- 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 technological development do not necessarily track overall GDP).
- 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 and global average potency

2.2 Transition Function

2.2.1 AI Potency

The AI potency of actor n at time t, p_t^n, updates as follows:

$$p_{t+1}^n = \begin{cases} p_t^n + f(p_t^n, {reg_{dom}}_t^n, reg_{int_t}) & \text{with probability } 1 - g(p_t^n, {reg_{dom}}_t^n, reg_{int_t}) \\ p_t^n + m^{p_t^n} & \text{with probability } g(p_t^n, {reg_{dom}}_t^n, reg_{int_t}) \\ \sqrt[k]{\frac{1}{N} \sum_{i \in N} p_t^i} & \text{if } p_t^n < \sqrt[k]{\frac{1}{N} \sum_{i \in N} p_t^i} \end{cases}$$

Where:

- m, k are constants to be determined by the user, with $m, k > 1$
- f, g are parametrized functions that can be determined by the user, dependent on their beliefs. Some example options are given below.

f-Options:

- AI Growth: logarithmic, linear, exponential
- Regulatory Dampening: single-scale (international and domestic regulation are on a single scale. The dampening effect is the greatest of the two), combined force (international and domestic regulation both have separate dampening effects)

> For example, if f represents exponential growth with single-scale regulatory dampening:
>
> $$f = \frac{1}{\max(reg_{dom_t}^n, reg_{int_t})} ep_t^n$$

g-Options:

- Progress: global average, progress independent, actor progress dependent
- Regulation: single-scale, combined force, regulation independent (regulation has no impact on new unthought-of areas of research)

> For example, a regulation-independent global average g could be represented by:
>
> $$0.01 \cdot 1 \cdot \frac{1}{N} \sum_{i \in N} p_t^n$$

2.2.2 Welfare

Welfare represents civilians' well-being and is a variable that exists only for countries. While creating a mathematical model of human well-being is reductive and imperfect, we attempt to give an overview of the general trends with regard to AI progress using the following equation for the welfare of a state n at time t, w_t^n:

$$w_t^n = w_{t-1}^n + r + \frac{1}{10} \ln\left(\frac{p_t^n}{p_{t-1}^n}\right) \cdot \mathcal{U}(0,1) - X_{glob_t} - X_{loct}$$

Where:

$\mathbf{w_{t-1}^n}$ is the welfare at the previous time-step, as we assume a degree of continuity over time.

$\mathbf{r}$ represents countries' ability to recover after a disaster occurs. Welfare recovers up to the previously reached maximum welfare.

$\frac{\mathbf{1}}{\mathbf{10}} \ln\left(\frac{\mathbf{p_t^n}}{\mathbf{p_{t-1}^n}}\right) \cdot \mathcal{U}(\mathbf{0}, \mathbf{1})$ represents the possible gains from AI development, such as improved medicine and teaching, with $\mathcal{U}$ as the uniform distribution.

$\mathbf{X_{glob_t}}$ represents possible global harms, i.e. those that are dependent on global progress and where global welfare is affected. For example, these could occur due to bioweapons or nuclear harm.

$\mathbf{X_{loct}}$ represents possible localised harms, i.e. those that occur in one country and are dependent on local AI potency and legislation. In general, these would be harms that occur legally and using AI available in that country, such as social monitoring or psychological manipulation.

There are also risks that are dependent on global AI potency but may be localised in their effect such as cyber- or conventional weapon attacks by terrorist organisations. However, in our increasingly connected world a cyber- or bioattack in one country is likely to have severe effects on other countries as well. Hence, for simplicity, we do not include this category in our model.

Recovery A country's speed of recovery after disaster varies depending on a country's resilience and the type of disaster that has occurred. However, in a large number (though by no means all) of cases initial recovery happens quickly then slows over time. We represent this slow-down effect using the difference between the recovery maximum and current welfare over time, divided by a user-decided factor, k to affect speed. To capture the variability in disaster effects, the maximum recovery that a country can achieve after a disaster (if no new improvements are achieved) is their previous maximum welfare multiplied by a factor between a and b, $a, b > 0$, chosen according to the uniform distribution. For example, suppose $a = 0.5$ and $b = 1.25$, this would represent the fact that disasters can permanently lower a nation's well-being without new developments, however it also leaves room for the possibility that a country might rebuild better than before. Hence, we have:

$$r = \begin{cases} \frac{\texttt{recovery\_max} - w^n_{t-1}}{k} & \text{if } w^n_{t-1} < \texttt{recovery\_max} \\ 0 & \text{otherwise} \end{cases},$$

where $\texttt{recovery\_max} = \max_{q \in T}(w^n_q) \cdot \mathcal{U}(a, b)$, with $\mathcal{U}$ the normal distribution.

Global Harms Global harm represents the risks from large-scale AI-enabled terrorist attacks, such as the creation of bioweapons, a loss of control scenario where a powerful autonomous AI system acts contrary to its intended design and over which control cannot be regained, or a nuclear conflict fuelled by AI-enabled algorithmic relation. This could cause direct harm to individuals and infrastructure or lead to unintended international conflict.

We model global harm by sampling from two distributions. First, to determine whether a global harm event will occur, we perform a Bernoulli trial, then, to determine the scale of that event, we sample an exponential distribution.

We model the Bernoulli trial with success probability $p = m \cdot (1 - \frac{reg_{int_t}}{max_{i \in N} \log_1 5p^i_t})$. Thus, as international regulation increases, the likelihood of an AI-caused globally harmful event decreases and as the maximum level of AI potency in the world increases, the likelihood of such an event increases. We have chosen to use maximum AI potency to reflect the real-world phenomenon that AI programs are used across jurisdictions and, while greater controls through global regulation lower this risk, organisations seeking to perform terrorist attacks or an out-of-control AI will be able to circumnavigate them.

We model the exponential distribution as $\lambda = \frac{1}{\log_5 max_{i \in N} log_5 p^i_t}$. As maximum AI potency increases, the maximum possible harm increases and the probability density becomes less focused at low harm and more evenly distributed across harm levels.

Local Harms Local harms represent either legal or accidental harms, such as erosion of local culture, technology-enabled discrimination, psychological manipulation, or a localised loss of control scenario. Their probability of occurrence and harm level decrease with local regulation but increase with the level of potency reached in that country.

We again sample from two distributions to model local harms. They occur with $p = k \cdot (1 - reg^n_{dom\,t})$, and the harm level is sampled from the exponential distribution with $\lambda = \frac{reg^n_{dom\,t}}{\log_5(p^n_t)}$.

2.2.3 Domestic Regulation

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

If n is:

- A country, domestic regulation is updated as follows:

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

- A company, then reg^n_{dom} tracks the regulatory state of the country in which the company is domiciled.
- A terrorist organisation, then $reg_{dom\,t} = 0, \forall t$

2.2.4 Counter-proliferation

In the current version of the model, the counter-proliferation mechanism in place is limited to global 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'

2.2.5 Friendship

We model friendship with an $N \times N$ matrix for every time step t. For a pair of actors n, m:

1. Initialise $C = 0$
2. If $m : reg_{dom}$ then $C+ = k_1 \cdot ({reg_{dom}}_t^m)^2$
3. Update C according to the relevant international regulation combination in the following table:

N (base actor)	M (other actor)	Reaction	$C+=$
reg_{int}	reg_{int}	Good	k_1
reg_{int}	der_{int}	Bad	k_2
der_{int}	reg_{int}	Slightly Bad	k_3
der_{int}	der_{int}	Ok	k_4

   Where $k_1 > k_4 > k_3 > k_2$
4. $R[n, m]+ = C \cdot \frac{1}{2}(R[n, m] + 1)$

Note:

- For each of the above, k_i can be adjusted according to the user's preference, including negation.
- The update function (line 4) assumes that the degree of closeness between two states affects the degree to which the other states' actions change the relationship.

3 Preliminary Model

A preliminary version of the model has been completed, with the structure given in section 2. It can currently handle 100+ agents for 10000 steps in less than two minutes on an M1 Mac without parallelisation and optimisations. An initial version of the model, including Monte-Carlo simulations, analytic tools, and parallelisation will be published as a Python notebook in this GitHub repo in late 2024.

3.1 Preliminary Outputs

In the following examples, we use five agents, whose utility functions are drawn from the following:

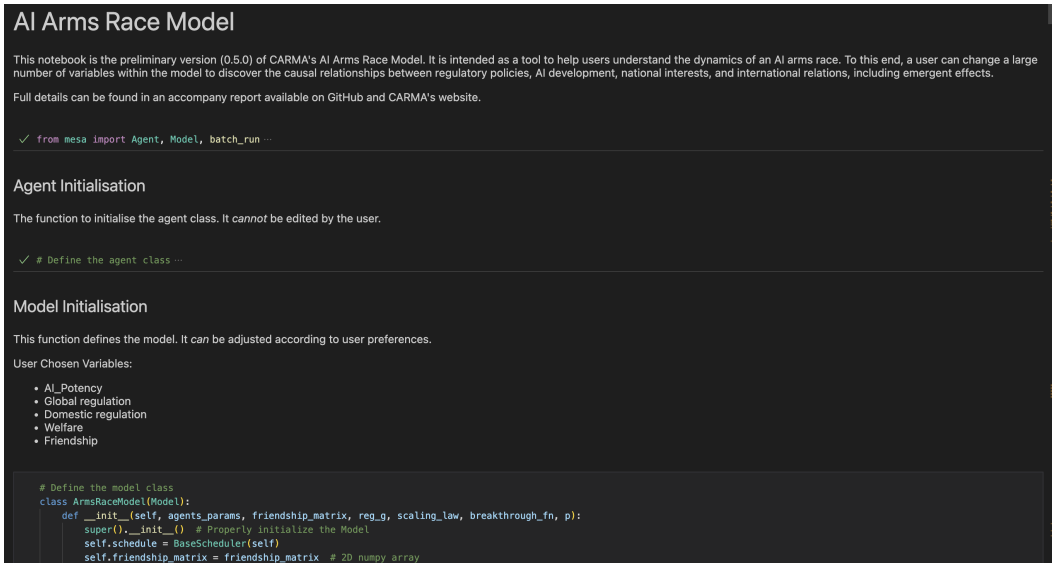
HMin , a harm minimizer, regulates domestically, pushes for international regulation, and does not develop

GRD pushes for global regulation but develops domestically. Does not change domestic regulation from its initialised state.

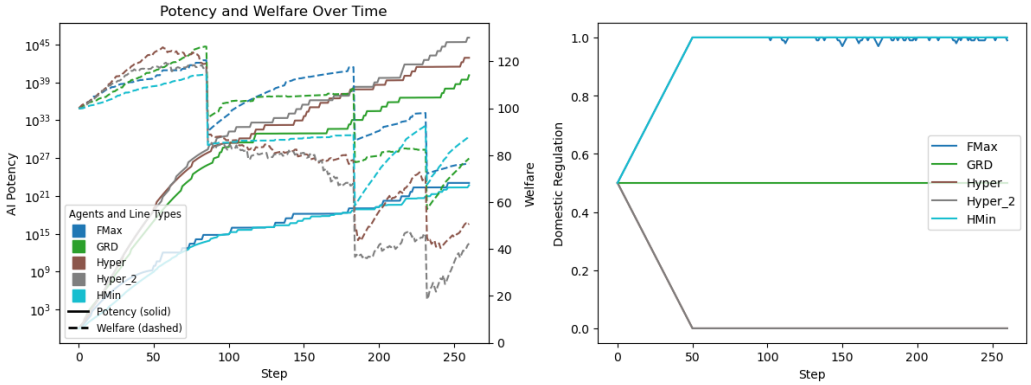
FMax , a friendship maximize, attempts to raise friendship as much as possible by matching the majority's choice on international regulation. Chooses other actions randomly.

Hyper prioritises scaling: always develops, always deregulates.

Figure 1: A preliminary version (0.5.0) of the model.



As initial settings, we use a 260-step simulation (equivalent to five years if each step is interpreted as a week). Each agent is initialised with $regulation_{dom} = 0.5$, $welfare = 100$ and $AI\_Potency = 1$, unless other specified. AI potency updates via the equation described in section 2.2.1, with $f(p_t^n, reg_{dom}{}_t^n, reg_{int}{}_t) = 2 \cdot p_t^n \cdot (1 - max(reg_{dom}{}_t, reg_{glob}{}_t))$ and $g(p_t^n, reg_{dom}{}_t^n, reg_{int}{}_t) = 4 \cdot p_t^n$. Global regulation is determined by majority vote and increases in increments of 0.01.



(a) A plot from the preliminary version of the model of AI potency (left axis, logarithmic) and welfare (right axis) against time.

(b) A plot of domestic regulation against time

Figure 2: A 260-step simulation of five agents where the scenario *does not* allow for a global moratorium on AI development.

At the beginning, until a year and a half in, all agents' welfare increases, with the

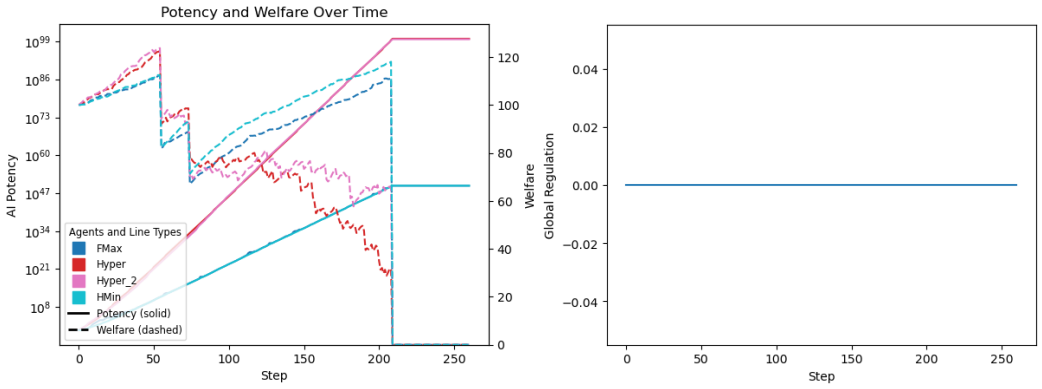
hyperscalers outperforming other agents. These advances in welfare would come through advances such as improved education, infrastructure, scientific and technological innovation. The hyperscalers' lack of local regulation means both that they develop at a faster rate than agents with safeguards in place and that they can integrate AI systems into society more quickly than other actors and hence reap the benefits more quickly as well. The friendship maximizer and harm minimizer quickly develop local regulation, although the friendship maximizer experiences some political back-and-forth on this over time. For the first year and a half, all agents experience some variation in welfare, including some local variance from issues such as widespread structural discrimination or cultural loss, however the general trend is clear. That is until a global disaster hits, a year and a half after the beginning of the model, affecting all countries. Such a disaster could come in the form of a terrorist bioweapons attack, AI arms-race-powered nuclear conflict, or mounting organizational risks taken with AI that eventually lead to disaster, as happened with Chernobyl and Three Mile Island. The harm minimizer and friendship maximizer recover from the disaster slowly and to differing levels. Their slower speed of development combined with local regulation encourages safe practices on a domestic level, but they are still vulnerable to these global shocks. The hyperscalers, on the other, hand struggle to recover, suffering from repeated local harms, such as psychological manipulation and localised AI-caused accidents. When the next global disaster hits a little over three years in, their welfare is at a significantly lower starting point than that of the other actors, despite their significantly higher levels of AI potency. This situation is repeated again a year later, with this disaster dropping one hyperscaler's welfare down to a fifth of its starting point. All nations begin to recover, but it is clear that a further local or global catastrophe could send them all over the edge.

Even for the harm-minimizing agent, we see in figure 2 that welfare drops significantly when a global harm occurs. Although it has not developed AI at all from its initial starting point, it is still subject to cross-jurisdictional global harms such as AI-caused pandemics and the increased risk of global conflict due to, for example, arms-race-fueled hot conflicts among the great powers. As AI potency continues to increase, the drops in welfare will become steeper, making even the harm-minimizing agent vulnerable to extinction. Due to these risks there is strong mutual vulnerability; it is in every country's national interest to not only slow development domestically but internationally as well. Whether this is achieved via global regulation or other counter-proliferation methods, any scenario where AI potency is unchecked in one nation leads to widespread harm for all nations.

Examining domestic regulation in figure 2b, we see that the two hyperscaler agents completely deregulated, which meant that they were more vulnerable to local shocks, with Hyper_2 nearly going extinct near step 240, and to global shocks, as they may be occurent with global shocks, as occurred near step 170. Further, we note that the 'Develop and Regulate Globally' actor achieves the same level of AI development as the hyperscaler, but due to higher domestic regulation, manages to avoid a sharp drop in welfare.

If we remove the agent pushing for global regulation (figure 3), there becomes a majority pushing for global deregulation and we see global AI potency become smoothly exponential and global AI risks in turn increase in their severity. While the friendship maximize (just) survives the first crash due to a combination of high local regulation and sheer luck, if AI potency continues to grow, reaching the levels of the hyperscalers, it will also cease to exist.

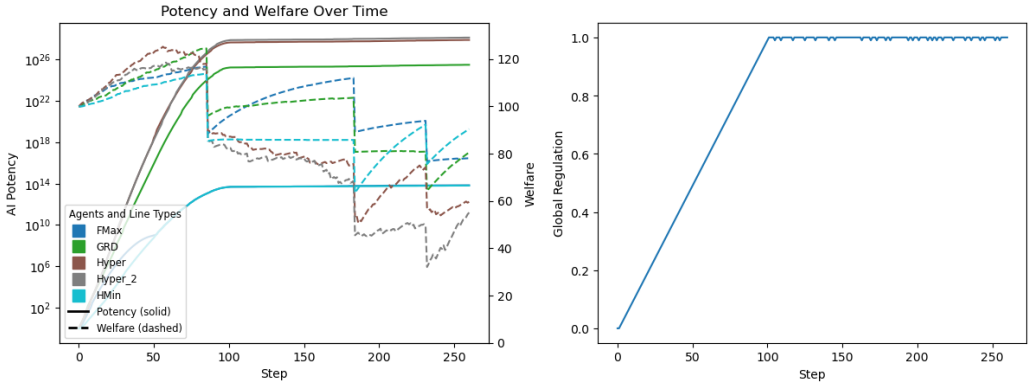
If we instead allow for a global moratorium on AI development (figure 4), where both normal development and breakthroughs cease at the maximum level of global regulation level, then we see in figure 4a that the global decrease in welfare from AI-induced shocks is significantly mitigated (note that the hyperscaler agent, which has no domestic regulation



(a) A plot of AI potency (left axis, logarithmic) and welfare (right axis) against time from the preliminary version of the model.

(b) A plot of global regulation against time

Figure 3: A 260-step simulation of four agents where the scenario *does not* allow for a global moratorium on AI development.



(a) A plot from the preliminary version of the model of AI potency (left axis, logarithmic) and welfare (right axis), against time.

(b) A plot of global regulation against time

Figure 4: A 260-step simulation of five agents where the scenario *allows* agents to achieve a global moratorium on AI development by voting for the maximum level of global AI regulation.

on AI use, still suffers significantly).

3.2 Preliminary Lessons

While the model is designed to show the effects of local and global harms in an AI arms race, the effects are striking.

AI Potency increases exponentially and is likely to very quickly, within five years (see figure 3), lead to human extinction in cases where there is no global regulation. The results

in section 3.1 were obtained using an AI-potency scaling function of $p_t^n = kp_{t-1}^n$, which leads to exponential growth over time but is not itself an exponential function. Thus, *even with relatively slow growth*, unregulated AI proves to be an existential risk.

Further, in each of the three figures 2, 3, and 4 above, the hyperscalers consistently outperform the other utility functions for the first 50-75 steps, or one to one and a half years in real terms. If one were within these scenarios, it would seem as if the hyperscalers were taking the right approach to AI legislation, potentially even encouraging other countries to deregulate to gain the same benefits. Similarly, we see that while the harm minimizer begins behind in each of the three figures, by the end in each of them (or before extinction in the case of 3), it has higher welfare than other agents. Both the the hyperscalers' and the harm minimizer's welfare are explained by the same underlying dynamic, explained in the next section, Key Insights.

4 Discussion of the Model

4.1 Key Insights

The military and economic superiorities that AI brings encourage an arms race where each nation sprints to stay ahead of the competition. This in turn discourages regulation as well as caution — anything which can slow AI development and its wider integration into society. This creates a situation where risks — whether local, such as economic displacement and state surveillance, or global and catastrophic, such as bioweapons, loss-of-control scenarios, and nuclear conflict — are magnified by the very structures designed to accelerate AI capabilities. In sum, AI's benefits and risks force nations' attitudes toward it to become a collective action problem.

Although the myriad pathways to AI risk are varied, there is an underlying structure to them. By examining the same problem through the lens of control theory and the analogy of financial leverage, we can gain a more nuanced understanding of the problems that humanity faces.

Control theory offers a structured framework for understanding how human oversight will struggle to keep pace as AI systems grow in speed, complexity, and impact. As AI potency increases, human faculties of data processing, decision-making, reaction times will increasingly diverge from AI's capacity for rapid, high-leverage actions. This creates a situation where humans begin to defer to AI systems because they can no longer understand and make the necessary small corrective adjustments quickly enough within a system of AIs. We can describe this situation using the metrics of control theory:

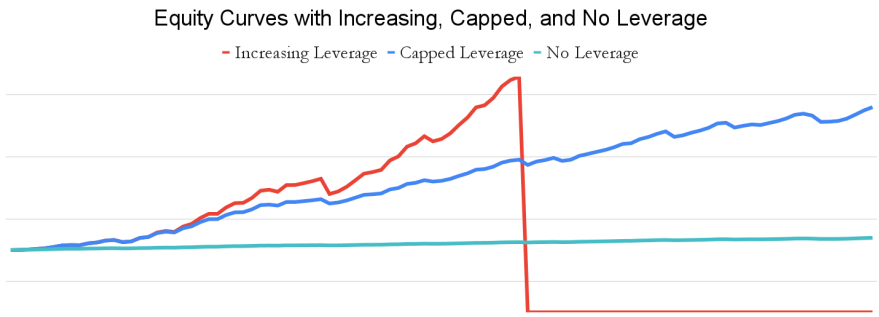
AI's higher *sampling rate* — the frequency with which a system updates its control actions — would leave humans incapable of keeping pace with rapid updates in real-time. AI's faster than human *response time* to changes in input means that introducing human oversight could also introduce dangerous delays. With *information processing capacity* and *throughput* — measuring how much data and how many decisions respectively an actor can handle per unit of time — the widening disparity between human processing power and AI efficiency becomes stark, reinforcing the risks of humans falling out of the control loop.

The growing differences in each of these metrics points towards the fact that as AI grows more powerful, the ability for humans to meaningfully intervene in high-speed decision-making will be severely diminished, increasing the risk of unintended behaviors or loss-of-control scenarios.

In each of the above examples, there is a large divergence between welfare and AI potency. To understand this phenomenon we can draw a parallel to a leveraged finance scenario where the leverage factor itself grows with time (see figure 5). We can construe AI Potency as a form of leverage, where increasing AI capability amplifies the instructions, positions, and biases communicated to it, as well as the potential benefits and harms of those over time. At the beginning, welfare, tracking AI potency, may follow an upward trajectory which could be mistaken for sustainable progress. However, this initial increase is due to luck, as the system becomes increasingly fragile over time. Just as highly leveraged traders may enjoy short-term success before eventually experiencing a catastrophic loss, the growth in AI potency similarly heightens the risk of severe, uncontrollable failures. The growing speed and complexity of a trader's control system would make our increasingly leveraged trader lose confidence in her ability to get ahead of the brunt of market moves with trades, and she instead simply accepts what her position and the market bring. This analogy shows why welfare and AI potency should be expected to diverge: unchecked increases in capability will eventually surpass our ability to mitigate risks.

A collective action problem is solved, in the end, by collective action. To gain the benefits of AI, while avoiding the various pitfalls along the way, global cooperation is necessary.

Figure 5: A graph showing equity curves with increasing, capped, and no leverage, as an analogy to different countries' attitudes towards AI potency. This scenario has 95% positive weeks, and no week with more than a 2.5% dip in the underlying asset – still ruinous when leverage grows too high.



4.2 Model Advantages

- **Highly modular and customizable** Everything from the utility functions to the scaling laws is customizable. This has two functions. Firstly, it allows users to make the model reflect their beliefs about AI scaling laws, international relations, etc. Secondly, it enables the sensitivity analysis discussed in section 5.1 which is necessary to understand the complex dynamics at play.
- **Usage versatility** The simulation platform can be leveraged for a variety of purposes, including evaluating the decision-making of AI agents, regulators and researchers alike and for a wargaming platform to pitch humans against AI and/or pre-baked utility functions (see section 5.2).
- **High level of detail** This model is more detailed than traditional ARD models.

4.3 Model Disadvantages

- **Assumption-reliance** A common weakness which is mitigated to an extent by clarity on assumptions and modularity, however it still stands. The model should not be interpreted as an accurate depiction of reality.

- **Insufficient strategic interaction** While the model in its current form is sufficient for evaluating decision-making in a context that is relatively isolated (interactions are only possible via global harms, global regulation, and friendship), one crucial missing component is strategic interaction between agents: counter-proliferation directly between states. This will be added in a future version (see section 5.2).

4.4 Assumptions and Limitations of the Current Model

1. **AI potency is currently on a single scale**; that is, it does not presently allow for different AI capabilities to progress at different speeds, it simply measures AI potency as a single number.

2. **Domestic/International regulations are currently each on a single scale**. The present model does not allow regulation to apply to different aspects of AI.

3. **The relative sizes of local and global harms is not well calibrated.** As applications of AI tend to proliferate quickly across the legitimate internet and the dark web, the ability for those applications to cause harm may exceed the control that local governance can exert. Global harms, on the other hand, include global catastrophic risks, which can quickly exceed the devastation wrought by local calamities. The same level of destruction wrought upon multiple countries results in worse outcomes for all than if it had only occurred in one. Consider, for example, global trade, on which large swathes of the global population are reliant; a disaster affecting multiple key economies would disrupt supply chains, leading to shortages, economic instability, and cascading failures across other sectors.

4. **Recovery is currently insufficiently grounded.** A country's ability to recover from a harmful event is greatly dependent on the type of event they have suffered, as well as the country's own resources and preparedness. Currently, these variations are not taken into account by the model, with it modelling the uncertainty only via a uniform distribution affecting the level of achievable recovery.

5. **The possibility of AI alignment is not yet implemented.** Increasing AI safety can be modelled via increasing regulation and welfare recovery between harmful incidents. However, the model does not presently include the possibility of reaching complete AI alignment and thereby totally countering the harmful effects of AI.

6. **We do not yet include an economic/GDP variable.** Instead, we include the dynamic of *increased* p_i $\rightarrow$ *increased speed of* p_i *growth*, which shortcuts the loop of better *AI* $\rightarrow$ *enhanced economic competitiveness* $\rightarrow$ *more resources for AI research* $\rightarrow$ *better AI*. Similarly, *increased* p_i $\rightarrow$ *increased dominance* can stand in for the general economic benefits and dominance of better AI. This assumes that AI will have economic benefits and that some of those will be directed towards further AI research

7. **We currently model e.g. the US as a single entity with a single domestic regulation-dampening factor.** This is inaccurate as the US has a federal system whereby AI regulation differs from state to state (e.g. see the recently proposed SB1047 in California). As most of the largest AI companies are based in California, we propose setting the 'initial' dampening factor and other relevant variables to those applicable to California rather than the US generally.

5 Coming Next

5.1 Analysis

To use this model as a tool for understanding the dynamics of an AI arms race, we must have a good understanding of its robustness and how its output varies with input. Hence we will set up a sampling base, each parameter setup run on several seeds for robustness, and then perform the following analyses, which we expect to introduce in the same release:

- **Sensitivity Analysis** to determine how changes in input parameters affect the outcomes. In our case, this allows us to determine which levers, in terms of regulation or countries' interests, need to be adjusted to mitigate an arms race scenario.
- **Uncertainty Analysis** to determine how uncertainty in the parameters influences variability in the output. This will provide a measure of the robustness of the model.
- **Policy Inference** on the sensitivity analysis to infer what policy or norm changes would best mitigate harmful race dynamics.

5.2 Future Model Features

Future versions of the model will contain, in order of priority:

1. **Monte Carlo Simulation** to give agents utility functions that are of the form 'maximise variable x'.
2. **Separation of Actors** into different categories such as countries, companies, and groups that engage in political violence. Actors will be differentiated along the following axes:
    - Action spaces
    - Degree of influence over regulation (implemented via a weighting factor)
    - Degree to which they follow regulation (e.g. terrorist organisations would not follow regulation)
3. **Counter-proliferation** whereby countries can choose to use some of their resources and affect their diplomatic relationships with other nations by choosing to inflict diplomatic, economic, or military sanctions on another country to slow, stop, or negate their AI potency .
4. **LLM-integration** allows agents to be controlled by an LLM agent instead of an expected-utility-maximising function.

5. **Wargaming** would enable a user to place themselves into the simulation, choosing one actor's actions in order to gain a better understanding of the model dynamics by gaining on-the-ground experience.

Transforming the world of international regulations and relations into a mathematical model will necessarily be a reductive simplification that misses much of the complexity of the real world. We expect to iterate on the transition function over time, improving its reflection of real-world dynamics. If you would like to help us improve the model, please contact the corresponding author. *All feedback will be gratefully received.*