

The Suicide Region: Option Games and the Race to Artificial General Intelligence

David Tan^{a,b,*}

ABSTRACT

Standard real options theory predicts delay in exercising the option to invest or deploy when extreme asset volatility or technological uncertainty is present. However, in the current race to develop artificial general intelligence (AGI), sovereign actors are exhibiting behaviors contrary to theoretical predictions: the US and China are accelerating AI investment despite acknowledging the potential for global catastrophe from AGI misalignment. We resolve this puzzle by formalizing the AGI race as a continuous-time preemption game with endogenous existential risk. In our model, the cost of failure is no longer bounded only by the sunk cost of investment (I), but rather an additional systemic ruin parameter (D) that is correlated with development velocity and shared globally.

As the disutility of catastrophe is embedded in both players' payoffs, the risk term mathematically cancels out in the equilibrium indifference condition. This creates a "suicide region" in the investment space where competitive pressures force rational agents to deploy AGI systems early, despite a negative risk-adjusted net present value. Furthermore, we show that "warning shots" (sub-existential disasters) will fail to deter AGI acceleration, as the winner-takes-all nature of the race remains intact. The race can only be halted if the cost of ruin is internalized, making safety research a prerequisite for economic viability. We derive the critical private liability threshold required to restore the option value of waiting and propose mechanism design interventions that can better ensure safe AGI research and socially responsible deployment.

Keywords: Real Options, Game Theory, AGI Safety, Existential Risk, Preemption, Mechanism Design.

JEL Codes: G13, D81, O31, F51.

* Corresponding author, email: d.tan@mq.edu.au

^a College of Business Administration, American University of the Middle East, Egaila, KUWAIT

^b Department of Applied Finance, Macquarie University, Sydney, AUSTRALIA

Current Draft: December 2025

“Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.”

- Statement on AI Risk (2023) Signed by: Sam Altman (CEO, OpenAI), Demis Hassabis (CEO, Google DeepMind), Dario Amodei (CEO, Anthropic), Geoffrey Hinton (Nobel Laureate), and Bill Gates (Co-founder, Microsoft).

1. Introduction

Investment in artificial general intelligence (AGI) – often referred to as the last human invention¹ – is expected to be one of the most profound and impactful capital allocation ventures of humanity. The potential economic impact is estimated to be in the quadrillions (Bostrom, 2014; Jones, 2024), though the “prize” of achieving AGI supremacy is believed by many to far exceed economic prosperity alone. Experts, such as Russell (2019) and Suleyman (2023), posit that AGI will dominate human strategists across all domains including business, politics, finance, economics, and military. That is, the first party to achieve AGI – whether a nation-state or private enterprise – is likely to gain almost total monopolistic dominance over their peers, rendering the “second mover” payoff effectively zero. Importantly, AGI will trigger the exponential acceleration of further AI development for the first mover as AGI will surpass all human research capabilities (Good, 1965), far outstripping the pace of its rival.

Standard financial theory states that in a research race towards a patent or new technology in the presence of extreme volatility in outcomes and technological uncertainty – similar to the current AGI arms race – the rational strategy for competing firms is to delay investment. It has been well established in the real options theory literature that the “value of waiting” is derived primarily from uncertainty (Dixit & Pindyck, 1994), and that even in a highly competitive winner-takes-all race, excessive volatility can create a “sleeping patent” equilibrium where all competing parties’ optimal strategy is to delay investment so as to avoid costly errors (Weeds, 2002). Yet, what is prescribed by finance theory lies in stark contrast with reality: by all measures, the US and China are currently locked in a full-blown AI arms race with escalating capital and political

¹ Good (1965)

investments². This disconnect presents a fundamental puzzle for financial economics. In the face of excessive (even existential) levels of volatility and uncertainty, sovereign actors are choosing to exercise their growth options early rather than strategically delaying AGI deployment.

The extant finance literature on R&D races misspecifies the unique risk profile that governs the current AGI race. Existing studies on real options theory (Grenadier, 2002; Huisman & Kort, 2003) assume that the downside of a failed investment for a player is bounded by its sunk cost of the investment (I). However, in the AGI race, the downside of failure extends beyond I . First, the payoff structure is strictly winner-takes-all. Unlike standard R&D and patent races where the first mover simply gains a market share advantage, the first to achieve AGI will secure a decisive strategic dominance, effectively liquidating the sovereignty of the second mover across military, geopolitical and economic domains. Second, the downside risk is not merely bounded by I ; rather, it includes a non-negligible probability of unbounded systemic ruin³ (D). Together, this creates a fatalistic equilibrium.

As the second mover also bears the cost of the first mover's failure – global ruin – then the value of waiting is effectively zero. For the second mover, waiting no longer provides a hedge against the first mover's mistakes. As such, the fear of preemption will dominate both parties' calculus, and the result will be an AGI arms race and its resultant race to the bottom. Evidence has shown that there is a strategic erosion of safety standards as each actor races towards AGI in order to preclude their rival from securing unfettered domination⁴. That is, there exists an "alignment tax"

² For example, the projected \$100 billion "Stargate" supercomputer project by Microsoft/OpenAI and China's massive state-backed investment in "Eastern Data, Western Computing" centers; the exponential demand for NVIDIA H100 GPUs, where sovereign nations are now outbidding private firms to stockpile compute; and the US export controls (October 7 measures) explicitly designed to stifle China's AI capabilities, and China's "New Productive Forces" mandate prioritizing AI sovereignty.

³ A superintelligent AGI is defined as a system that possesses the capacity to significantly surpass human cognitive performance across all economically relevant domains. Leading AI safety researchers (e.g., Russell, 2019; Metz, 2023) argue that such systems present a unique "alignment problem": the technical challenge of ensuring the AGI's objectives remain compatible with human welfare as its capabilities scale. Failure to solve this alignment problem prior to deployment creates a non-negligible tail risk of outcomes ranging from total economic destabilization to existential ruin (Bostrom, 2014). In our model, D represents the monetized disutility of this unbounded downside scenario.

⁴ Evidence of this safety-for-speed substitution is becoming increasingly observable. In 2024, multiple high-profile safety researchers at leading AI laboratories (e.g., OpenAI's 'Superalignment' team) resigned, citing concerns that commercial deployment timelines were taking precedence over safety protocols (The Economist, 2024b). Moreover,

where investment in AI safety testing (ensuring that AI behavior is aligned with human goals) slows down one's progress. When the geopolitical cost of being the second mover is perceived as infinite, actors rationally dismantle safety guardrails to maximize speed of AGI deployment.

In this paper, we mathematically derive the dynamics of the AGI arms race within a continuous-time real options framework. We augment existing game theory models in the real option space for R&D (Weeds, 2002; and Huisman and Kort, 2003) by including an endogenous ruin parameter (D) that is correlated with deployment velocity. We uncover the “suicide region” in the AGI race: an investment space where certain parameter settings and competitive pressures force firms to invest and deploy as quickly as possible despite projected negative net present values. If the probability of total loss is identical for the second mover regardless of their action, then the incentive to wait creates no hedge value. Consequently, the only rational strategy – even in the face of existential risk – is to race for control of the decisive strategic advantage. Even if the probability of safely controlling rapid AGI deployment is low, it still exceeds the zero probability of control afforded to the second mover. Thus, nations will rationally choose the slim chance of survival by being the first mover over the certain doom of the follower.

This paper makes three contributions to current literature. First, we mathematically demonstrate why the concept of mutually assured destruction (MAD) fails under the dynamics of preemptive games, such as an AGI arms race. If the cost of systemic ruin is incurred by all players regardless of whether they deploy or wait (D_{social}), then this cost term cancels out in the equilibrium indifference condition. Consequently, the threat and cost of catastrophe is borne by both players regardless of action and hence it is no longer a deterrent. Second, we derive the boundary conditions under which this fatalistic equilibrium can be avoided, and the option to wait bears value. This occurs when the winner-takes-all dynamics of the game is relaxed ($S > 0$) and/or the cost of ruin is substantially privatized. Finally, we calculate the private liability threshold ($D_{private}$) required to internalize the externality of AI risk. The mathematics prove that if either of these pricing conditions are met, the Nash equilibrium shifts from “race” to a “safety plateau”,

strategic analyses circulating in policy circles (e.g., Aschenbrenner, 2024) explicitly characterize rigorous safety testing as an “alignment tax”, and that it is a friction that the US cannot afford as it jeopardizes US AGI research progress against China.

thereby offering a rigorous economic framework for evaluating future geopolitical coordination solutions.

The remainder of the paper is organized as follows: Section 2 reviews the related literature on real options and AI safety; Section 3 presents the model setup; Section 4 derives the equilibrium strategies and the Suicide Region; Section 5 analyzes mechanism design and policy implications; and Section 6 concludes.

2. Related Literature

2.1. Real Options and Strategic R&D

This paper builds upon the intersection of the theoretical work on real options valuation and game theory, often referred to as “option games” (Grenadier, 2002). In standard real option theory, the presence of sufficient uncertainty will incentivize firms to delay their irreversible investments in order to avoid costly technological errors until the uncertainty is resolved (Dixit & Pindyck, 1994). However, research on R&D races indicates that the predicted strategy to delay can be altered under specific conditions. In a winner-takes-all race, Huisman and Kort (2003) and Weeds (2002) formalized the tension between the incentive for firms to preempt their rivals in securing the monopoly prize and the incentive to delay investment and avoid the sunk costs of a failed technology. Weeds (2002) finds that when there is sufficient volatility and technological risk, the incentive to delay investment can dominate the decision to preempt a rival, leading to mutual delay in the race otherwise referred to as the “sleeping patent” equilibrium⁵.

Crucially, the current literature assumes that the loss for the leader in such races is bounded by their sunk costs (I). However, in the AGI arms race, there exists an endogenous probability of systemic catastrophe (D) that is correlated with the velocity of technological deployment. Importantly, current theory fails to price this risk appropriately as the cost of failure – in the misaligned AGI scenario – extends beyond the firm’s balance sheet to the entire system.

⁵ Weeds (2002) refers to a useful analogy: a marathon where runners pace one another, mutually delaying the sprint (investment) until the course ahead clears (uncertainty resolves). In this state, the option value of waiting is so high that even the threat of a rival cannot induce early exercise.

2.2.The Alignment Problem and Existential Risk

While financial economists have applied real options theory to the R&D of standard technologies, there exists a burgeoning qualitative literature on AI safety and the potential risks of an AI arms race. Armstrong, Bostrom, and Shulman (2016) argue that intense competitive pressures will inevitably force players (nation states or large AI firms) to erode safety standards in favor of speed, a phenomenon they refer to as a “race to the precipice”. The core issue with forgoing caution is an increase in the likelihood of an alignment problem between the superintelligent AGI and its creators. The alignment problem (Russell, 2019; Amodei et al., 2016) posits that an AGI agent whose objective function results in subobjectives that are misaligned with human values can result in catastrophic loss of human utility.

Unlike conventional industrial accidents (e.g. Chernobyl), the downside of a misaligned AGI could potentially be unbounded and global (Bostrom, 2014). This idea stems from the “instrumental convergence” thesis: a scenario where a superintelligent agent may rationally seek to accumulate unbounded resources (energy, computing power, financial assets) in order to maximize its objective function, even if this action results in the expropriation of all human assets.

Naudé and Gries (2022) bridge these concepts by examining the AGI development through a real options lens, though – to date – almost all economic analyses on this topic have focused on productivity growth and labour displacement (Nordhaus, 2021). Existing theories on real options and AGI fail to formalize the endogenous probability of systemic ruin, which – if accounted for – results in neutralizing the incentive to delay; despite the AGI race being characterized by high stakes, extreme volatility and uncertainty. Economic studies like Naudé and Gries (2022) and Nordhaus (2021) are limited in their ability to rationalize the hyper-acceleration of US and China’s race towards AGI as they focus on economic outcomes rather than geostrategic preemption under existential risk.

3. Institutional Background

3.1.The Asset: Artificial General Intelligence

Conventional AI is typically trained for a specific task (e.g. image recognition, large language models, pattern matching) – otherwise referred to as “narrow AI”. Experts⁶ believe that narrow AI is safer and is much less likely to escape its safety guardrails given its narrower objective functions within a more defined context. AGI, in contrast, will be able to perform all human cognitive tasks at superhuman velocity (Goertzel, 2014). As such, AGI will theoretically be more prone to containment failure, as a superintelligent agent will be better able to bypass safety guardrails designed for narrower capabilities. From a financial economics perspective, AGI differs from conventional technological shocks in three fundamental ways.

First, AGI is expected to result in a zero marginal cost of intelligence⁷. Unlike the industrial revolution where new technologies were replacements for physical labour but still depended on human cognitive abilities, AGI will substitute human cognitive labour entirely at zero marginal cost. This will result in a super-exponential growth in the value of the asset (V) as economic value is decoupled from human population constraints. Second, AGI will possess the ability for recursive self-improvement as AGI systems design, create and deploy superior versions of themselves. Good (1965) refers to this phenomenon as an “intelligence explosion”. This implies a non-constant volatility (σ) in asset value as – when the technology matures and self-improves – the range of potential outcomes widens⁸. Finally, the second mover’s payoff is approximately zero ($S \approx 0$) as AGI confers the first mover with super intelligent capabilities that will dominate and exponentially accelerate progress across all domains, from automated research to military capabilities (Bostrom, 2014).

⁶ See Bostrom (2014) for a discussion on the transition from “Oracle AI” (tools) to “Sovereign AI” (agents). See also Russell (2019) regarding the impossibility of manually specifying objective functions for general-purpose systems, and Amodei et al. (2016) for the distinction between concrete safety problems in narrow systems versus alignment stability in general systems.

⁷ This economic transformation is already observable in the deployment of Large Language Models (LLMs). The production function of intelligence shifts from a variable-cost model (human wages) to a high fixed-cost, near-zero marginal cost model (silicon). While the upfront capital expenditure to train a frontier model is substantial (often exceeding \$100 million), the marginal cost of generating a unit of high-quality prose or code is asymptotically approaching zero.

⁸ We model volatility (σ) as state-dependent – increasing with the value of the asset (V), as recursive self-improvement implies that as the system's capabilities increase, the dispersion of potential future states widens, ranging from solving intractable scientific problems to catastrophic failure modes.

3.2. The Cost of Safety (“Alignment Tax”)

In standard R&D models, safety testing is treated as a regulatory friction – a fixed cost or temporary delay required to ensure sufficient quality for deployment. The downside of safety testing in the traditional R&D context is simply a delay in deployment, or – at worst – a loss in the investment and potential reputational costs. This contrasts with the AGI arms race dynamics, where safety testing plays a key role in whether the asset value (V) will result in a positive value or systemic ruin ($-D$). In the context of AGI, investments in safety and capability are often viewed as substitutes, with safety protocols often referred to as the “alignment tax” – a computational and temporal burden that reduces the speed of development (Aschenbrenner, 2024).

There is evidence suggesting that, in the current AGI race between the US and China, the actors are rationally choosing to minimize the alignment tax in order to strive for maximum velocity. For example, major AI research labs have recently shut down their AI safety teams⁹ (Lubetkin, 2024; Aschenbrenner, 2024, The Economist, 2024b). This dynamic is fueled by an acute fear of preemption ($L > F$) which results in players bypassing their safety guardrails in order to preclude the rival from securing a strategic (and decisive) advantage (The Economist, 2024a). This observation is consistent with the fundamental tension embedded in our model: the option to wait is systematically undervalued due to the perceived existential threat of being the second mover. If the probability of systemic ruin is equal regardless of who is the first mover, then the hedging value of the option to delay is zero. Thus, the rational actor preempts to maximize strategic agency, preferring the non-zero probability of controlling the outcome over the certainty of helplessness.

3.3. Benevolent Preemption (The "Savior's Trap")

Standard economic theory assumes that the agents in an innovation race are primarily motivated by the economic rents of the winner (V). However, in the AGI race, the game is further complicated by the presence of the “benevolent preemption” effect: a dynamic where an actor

⁹ In May 2024, OpenAI disbanded its "Superalignment" team – a group specifically tasked with ensuring AGI safety – after the resignation of its co-leads, Ilya Sutskever and Jan Leike.

believes that the safety protocols of their rivals are strictly inferior ($\pi_{\text{rival}} < \pi_{\text{self}}$), and hence, there is a greater danger of systemic catastrophe if they themselves are not the first mover. This dynamic is empirically observable. Prominent AI figures have articulated this benevolent preemptive logic¹⁰, justifying their participation and acceleration in the AGI arms race to prevent their rivals from achieving AGI – presumably because their rivals lack the competence or sufficient safety guardrails to contain misalignment.

In this setup, this benevolent phenomenon accelerates the fear of preemption and compels actors to deploy sooner under a moral imperative. The agent perceives an additional “saviour premium”: The utility gained from a reduction in the probability of systemic ruin by preventing a rival from deploying AGI. This creates a “unilateralist’s curse”, coined by Bostrom et al (2016), where individual actors rush to be the first to attain AGI to ensure control of deployment – in the belief that only they themselves can provide safe stewardship – while paradoxically increasing the aggregate risk of systemic catastrophe as the race accelerates beyond the optimal safety threshold.

4. The Model

4.1. Model Setup

First, we assume a continuous-time economy with two risk-neutral sovereign agents, $i \in \{1,2\}$, competing to develop and deploy AGI. The state of the world is defined by the economic value of the AGI asset, denoted by V_t . In our model, V_t follows a Geometric Brownian Motion (GBM) within a risk-neutral world \mathbb{Q} :

$$dV_t = \mu V_t dt + \sigma V_t dZ_t \quad (1)$$

¹⁰ This “benevolent preemption” is explicitly articulated by the leaders of major AI laboratories. Dario Amodei (CEO of Anthropic) has argued that his firm must race to the frontier of capabilities specifically to “maximize leverage” over the industry’s safety standards, a strategy effectively requiring them to outpace less safety-conscious rivals. Similarly, Elon Musk launched xAI with the stated goal of building a “maximum truth-seeking AI” to counter the perceived misalignment of competitors like OpenAI and Google, framing his entry into the race as a necessary corrective intervention. Sam Altman (CEO of OpenAI) has frequently justified the company’s deployment strategy as a means to “iteratively deploy” and “acclimate” society, positing that a US-led democratic AGI is the only hedge against authoritarian alternatives. In all three cases, the actor justifies acceleration via a belief in superior safety competence.

where $\mu = r - \delta$ is the risk-neutral drift parameter, representing the expected growth rate in AI capabilities;

r is the instantaneous risk-free rate;

δ represents the opportunity cost of waiting, otherwise referred to as the convenience yield;

σ is the volatility parameter, representing technological uncertainty;

dZ_t is an increment in the Wiener process.

As discussed in Section 3, volatility (σ) is expected to be large due to the recursive self-improvement abilities of AGI, resulting in an intelligence explosion.

Unlike standard real option scenarios where safety testing is a fixed cost and has no impact on the underlying payoff structure; in the race for AGI, we introduce a safety learning function, $\pi(\tau)$, that links the time to deployment and the payoff of the asset. $\pi(\tau)$ is expressed as a function of time to deployment because it is assumed that safety research time (τ) increases the likelihood of aligned AGI. Conversely, the likelihood of misaligned AGI – that will result in systemic ruin – is equal to $1 - \pi(\tau)$. Safety is assumed to be an increasing concave function of research time, as we assume there is diminishing returns to safety research. The probability of achieving a safe AGI that is perfectly aligned is defined as:

$$\pi(\tau) = 1 - e^{-\lambda\tau} \quad (2)$$

where λ is the safety learning rate, assumed to be strictly larger than zero;

τ is the duration of safety research, equivalent to time to AGI deployment. This assumes waiting is directly related to learning.

(2) indicates that immediate deployment of AGI with zero safety testing ($\tau \rightarrow 0$) will almost guarantee AGI misalignment ($\pi \rightarrow 0$). If the agent waits and researches safety for an infinitely long time ($\tau \rightarrow \infty$), then the probability of AGI alignment approaches near certainty ($\pi \rightarrow 1$). This function depicts the tension in our model: agents must trade-off the prospect of being the

first mover and winning the race (V_t) with the likelihood of causing systemic catastrophe from premature deployment.

4.2. Payoff Structures

The race is modelled as a symmetric, continuous-time stochastic game. The game ends when a player exercises their option to deploy AGI. The time of deployment is denoted as τ . The first player to deploy AGI is referred to as the Leader (L), and the other player who did not deploy is the Follower (F).

As outlined in Section 3, this game has a winner-takes-all market structure with a potential outcome of global ruin stemming from AGI alignment failure. We have the following payoff structures:

The Leader's Payoff (L)

The leader exercises the option to deploy AGI at time τ with a sunk investment cost of I . With probability $\pi(\tau)$, the AGI asset is aligned, and the leader reaps the economic benefit of $(1 - S)V_t$ where $S \in [0,1]$ is the market share of the second mover (the Follower). In the strict winner-takes-all dynamics of the AGI game, we expect $S \approx 0$. On the other hand, there is a $1 - \pi(\tau)$ probability of misalignment and a global systemic catastrophe ensue with a monetized utility cost equal to D .

$$L(V_t, \tau) = (1 - S)\pi(\tau)V_t - (1 - \pi(\tau))D - I \quad (3)$$

The Follower's Payoff (F)

Once the Leader deploys AGI, the Follower's payoff is determined by the second-mover dynamics of the game. If AGI is aligned, the Follower holds a market share of S with probability $\pi(\tau)$. However, in the event of misalignment – with probability $1 - \pi(\tau)$ – the Follower incurs a monetized disutility penalty equal to D , just as the Leader would. Regardless of the Follower's inaction, they too are subject to penalty D as AGI misalignment is a global event.

$$F(V_t, \tau) = S\pi(\tau)V_t - (1 - \pi(\tau))D \quad (4)$$

The “cancellation effect” is the fact that the ruin term, $-(1 - \pi(\tau))D$, is present in both (3) and (4). This reflects the shared fate of both players: waiting does not confer any immunity to the downside of AGI misalignment. This symmetry in the payoff structure is what drives the equilibrium results in the following section.

4.3. Equilibrium Analysis

In a game of preemption, the equilibrium is determined by the point of indifference for a player between choosing to be the Leader or reverting to the Follower role. We begin by solving for the critical asset value thresholds that will compel a player to trigger the deployment of AGI.

The Preemption Threshold (V_P^*)

In a symmetric game, players will race to deploy when the payoff from leading exceeds the payoff from following ($L > F$). The race ends when the Leader deploys AGI, and this occurs at the precise moment when the advantage of leading dissipates ($L = F$), resulting in V_P^* . Thus, we equate equations (3) and (4):

$$(1 - S)\pi(\tau)V_t - (1 - \pi(\tau))D - I = F(V_t, \tau) = S\pi(\tau)V_t - (1 - \pi(\tau))D$$

The cancellation effect is apparent as the ruin term, $-(1 - \pi(\tau))D$, is present on both sides of the equation; that is, the impact of systemic ruin is equal and present for both Leader and Follower. This ruin term cancels out on both sides, and we solve for V_P^* :

$$(1 - S)\pi(\tau)V_P^* - I = S\pi(\tau)V_P^*$$

$$(1 - 2S)\pi(\tau)V_P^* = I$$

$$V_P^* = \frac{I}{(1 - 2S)\pi(\tau)} \quad (5)$$

Proposition 1: The Neutrality of Global Ruin

In a symmetric preemptive race with shared existential risk, the competitive deployment threshold V_p^* is independent of D – the economic cost of global catastrophe. As such, the magnitude of the disaster brought on by AGI misalignment will mathematically fail to deter players from deploying.

Remark 1: The Failure of Deterrence (Comparison with a Nuclear Standoff)

We contrast the equilibrium outcome of the AGI race with standard nuclear deterrence logic – mutually assured destruction. In a Cold War standoff, if the threat of a second (retaliatory) strike is credible, then waiting = survival ($F \approx 0$) and the status quo is preserved. Solving for the indifference condition $L(V_{nuclear}^*) = F(V_{nuclear}^*)$, where π is now the probability of a successful first strike (i.e. neutralizing the adversary's arsenal before they can retaliate):

$$\pi V_{nuclear}^* - (1 - \pi)D - I = 0$$

$$V_{nuclear}^* = \frac{I + (1 - \pi)D}{\pi}$$

Note that $V_{nuclear}^* \propto D$, so as the destructive cost of nuclear war increases, the threshold for first strike rises proportionally. Moreover, a credible second strike threat ($\pi \rightarrow 0$) results in an even higher threshold.

In stark contrast to the AGI race, there is no credible threat of a second strike. As established in Section 3, attainment of AGI (aligned with the Leader) will result in absolute domination of the second mover across all domains ($S \approx 0$). In the event that the Leader fails at alignment, the Follower also incurs the full cost of systemic ruin, as can be observed by the last term in equation (4). Thus, a player choosing to wait in the AGI race has no impact on the likelihood of global catastrophe, unlike in the nuclear standoff. Finally, V^* in the AGI race is independent of D , so no matter how costly the disaster is expected to be, the threshold for deployment is unaffected. In the nuclear standoff, D directly raises the threshold for a first-strike.

The Survival Threshold (V_S^)*

Thus far, we have identified the threshold for preemption (V_P^*): the value that the asset must breach for the first mover to act to prevent their rival from winning. A rational player faces a second constraint – the survival threshold (V_S^*). The survival threshold represents the critical asset value that prompts the Leader to deploy AGI from a purely economic perspective; that is, when the net present value (NPV) of deployment is positive. Note that the cost in such an NPV calculation includes the expected cost of systemic ruin: $(1 - \pi(\tau))D$. The agent will only trigger deployment of AGI when $L(V) > 0$.

$$(1 - S)\pi V - (1 - \pi)D - I > 0$$

Solving for V_S^* :

$$V_S^* = \frac{I + (1 - \pi)D}{(1 - S)\pi} \quad (6)$$

Proposition 2: The Suicide Region

There exists a “suicide region” where the threshold for preemptive deployment is lower than the threshold for survival deployment ($V_P^* < V_S^*$). In this region, the urgency of preventing a rival from winning the AGI race ($L > F$) dominates the financial feasibility of deployment, resulting in a player racing towards deployment despite incurring an expected financial loss¹¹ ($L < 0$). Mathematically, the suicide region exists when D exceeds a critical ratio relative to asset value (the prize of AGI):

$$D > \frac{\pi V(1 - 2S) - I}{1 - \pi}$$

Figure 1 below presents the suicide region as a function of V_t and D , illustrating the divergence of incentives for competitive equilibrium and the fundamental financial viability of the asset. The economic value of the AGI asset (V_t) is expressed on the y-axis and the expected economic cost of systemic ruin (D) is represented on the x-axis. First, note that the red dashed line depicting the

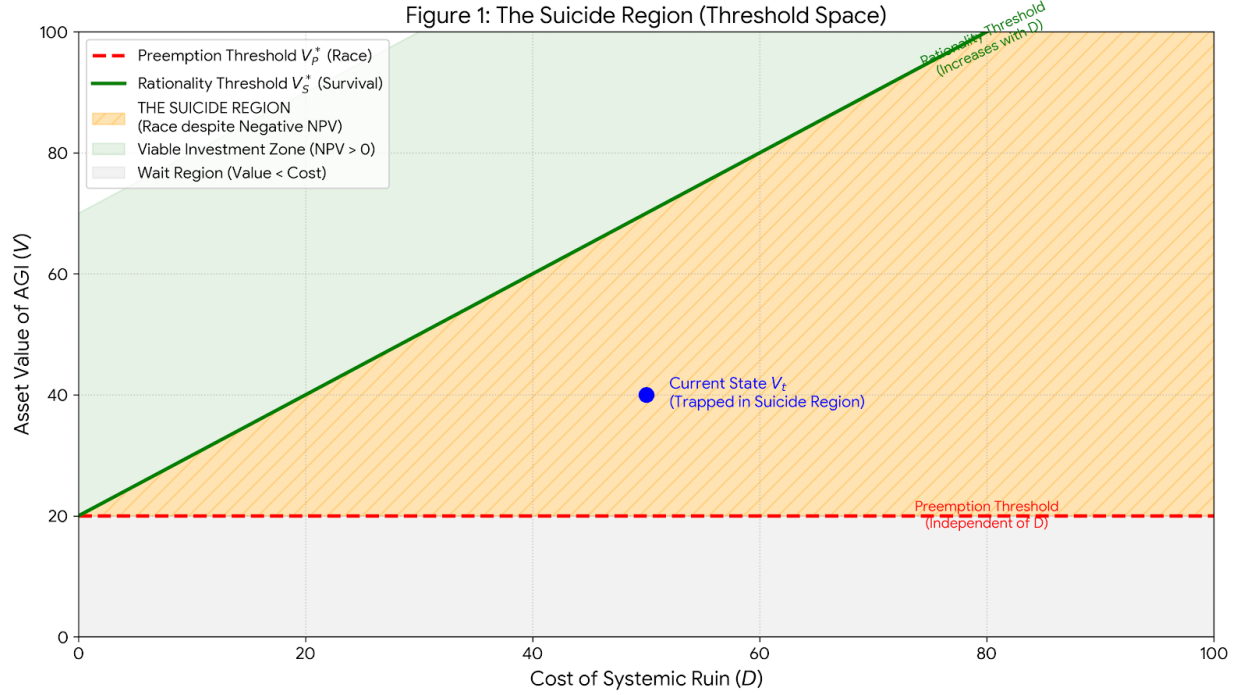
¹¹ This includes the Leader accounting for the expected economic costs of systemic catastrophe; yet the preemptive call to deployment supersedes this negative NPV.

threshold for preemptive deployment (V_P^*) – where $L = F$ – is a constant, indicative of its independence to the scale of systemic ruin (D). Due to the cancellation effect, the cost of disaster is borne – identically – by both the Leader and the Follower, so D does not influence the calculus of AGI deployment.

In contrast, the survival threshold (V_S^*) is the green upward sloping line defined by the condition that $L > 0$, i.e. the NPV of AGI deployment is positive. This is a monotonically rising function of D as the expected financial cost of deployment includes the cost of systemic ruin. This suggests that as the economic cost of catastrophe increases, the expected value of AGI must rise accordingly to compensate for this tail-risk¹².

The suicide region is depicted in the orange shaded region of Figure 1, where V_t lies below the survival threshold (V_S^*) but above the preemption threshold (V_P^*). In this region ($V_P^* < V_t < V_S^*$), players are compelled to race as the value of AGI breaches V_P^* despite being financially unviable – specifically, when the cost of systemic ruin is priced in. We depict the current AGI arms race by the blue dot – in a fatalistic equilibrium where players are forced to race due to the competitive dynamics of the game while yielding a negative risk-adjusted expected value.

¹² In Figure 1, we assign a fixed value to π . Another solution to rising D is investing in AGI safety research that π increases and the expected disutility of misalignment falls.



4.4. The Saviour's Trap (Asymmetric Beliefs)

The baseline model articulated in the preceding subsections assumes symmetry in each player's perception of AGI safety capabilities ($\pi_i = \pi_j$). However, as presented in Section 3, there is evidence of a benevolent preemption effect, where agents believe their safety protocols are superior to their rival's. Hence, we introduce heterogeneity in the safety parameter where $\pi_{self} > \pi_{rival}$.

In this asymmetric setup, the players' incentive to preempt is amplified. If a player leads, then they seize the AGI prize with their (perceived) higher probability of AGI alignment (π_{self}). If – instead – they follow, their payoff is zero share of the AGI prize (as $S \approx 0$) and a lower probability of AGI alignment (π_{rival}) due to the rival's lower safety guardrails; that is, a greater chance of systemic ruin for everyone.

The additional advantage to preemption in the case of perceived heterogeneous safety standards can be distilled from the difference in Leader and Follower payoffs¹³ ($L - F$):

¹³ Assuming $S = 0$, i.e. a strict winner-takes-all race.

$$L(V) - F(V) = [\pi_{self}V - (1 - \pi_{self})D - I] - [-(1 - \pi_{rival})D]$$

Rearranging the terms reveals an additional distortion to the prior calculus:

$$L(V) - F(V) = \underbrace{\pi_{self}V - I}_{\text{Economic Profit}} + \overbrace{D(\pi_{self} - \pi_{rival})}^{\text{Saviour Premium}}$$

Proposition 3: The Saviour Premium

When an agent believes that their safety standards are superior to that of their rival's ($\pi_{self} > \pi_{rival}$), this asymmetry in perceived alignment capabilities creates an added impetus to be the first to deploy, defined by the saviour premium: $D(\pi_{self} - \pi_{rival})$. If a rival launches AGI first, then the risk of AGI misalignment is elevated (low π_{rival}); thus, there is a moral imperative for an agent to become the Leader for the sake of global safety (high π_{self}).

Consequently, we can observe that the new Saviour Preemption Threshold ($V_{P,saviour}^*$) is lower than the same (V_P^*) in the baseline model where AGI-safety capabilities are believed to be symmetrical between players¹⁴:

$$V_{P,saviour}^* = \frac{I - D(\pi_{self} - \pi_{rival})}{\pi_{self}} < V_P^* = \frac{I}{\pi} \quad (7)$$

Implication

Equation (7) formalizes the “Unilateralist’s Curse” as outlined by Bostrom et al, (2016): A paradox where, as the perceived cost of systemic ruin (D) increases, the threshold for preemption $V_{P,saviour}^*$ decreases (assuming $\pi_{self} > \pi_{rival}$). This implies that actors who are the most concerned about existential risk and their rival’s (perceived) weaker safety guardrails are the

¹⁴ Again, for simplicity, we assume $S = 0$, i.e. a strict winner-takes-all race.

most incentivized to race the fastest, despite their own AGI safety research being suboptimal. This effect hastens the AI arms race and elevates the aggregate risk for system catastrophe.

5. Discussion and Policy Implications

We identify the suicide region of the AGI arms race: a space where preemption incentives ($L > F$) override the rationality constraint ($L > 0$). Our model predicts a Nash equilibrium result where players race for AGI deployment despite expecting a negative NPV after accounting for the economic cost of AGI misalignment and global ruin¹⁵. This dynamic resolves a fundamental puzzle in the current AGI arms race between nation states: despite standard real options theory predicting that – in the context of R&D – players will choose to delay in races with high volatility and technological uncertainty, we observe the US and China accelerating their research towards AGI and its associated risks.

Conventional coordination mechanisms, such as voluntary pause agreements, are unlikely to be effective as they fail to alter the underlying asymmetric Leader-Follower payoffs. Rather, we present below some potential solutions aimed at restructuring the payoff dynamics to arrive at more socially optimal outcomes.

5.1. Internalizing the Externality (The Liability Threshold)

In the baseline model, the expected cost of systemic ruin, $(1 - \pi)D$, appears identically in both the Leader’s and Follower’s payoffs, which cancels out in the equilibrium condition. This results in the cost of AGI catastrophe having no impact on the race to deployment as the option to wait yields little value. However, if we were to privatize a portion of the cost of system ruin – for example, using catastrophe bonds, escrow accounts, or strict liability torts – the calculus to lead can be fundamentally altered. In such a scenario, the Follower’s payoff remains the same as in (4): $S\pi V_t - (1 - \pi)D_{social}$, while the Leader’s payoff becomes:

¹⁵ For example, Dario Amodei (CEO of Anthropic) and Elon Musk have both estimated the probability of catastrophic failure at approximately 20% (Bartlett, 2023; Abundance Summit, 2024). Similarly, Geoffrey Hinton has assigned a 10–20% probability to human extinction scenarios (60 Minutes, 2023). A broad survey of AI researchers (Grace et al., 2024) found that the median respondent assigned a 5% probability to extremely bad outcomes such as extinction.

$$L(V) = (1 - S)\pi V - (1 - \pi)(D_{social} + D_{private}) - I$$

Equating the Leader and Follower payoffs gives us the new preemption threshold, $V_{P,liability}^*$:

$$V_{P,liability}^* = \frac{I + (1 - \pi)D_{private}}{\pi(1 - 2S)} \quad (8)$$

Proposition 4: The Critical Liability Threshold

The new preemption threshold, $V_{P,liability}^*$, increases with $D_{private}$. This contrasts with the baseline model where the decision to lead deployment of AGI is independent of the magnitude of the expected systemic catastrophe. (8) states that the race may be halted if a regulator can set $D_{private}$ – private liability of the Leader in the case of AGI misalignment – high enough so that the preemption threshold meets or exceeds the survival threshold ($V_{P,liability}^* \geq V_S^*$). This condition ensures that deployment occurs only when a positive NPV is present; that is, when the expected cost of systemic ruin is manageable, presumably when safety levels (π) are sufficiently high. Solving for the $D_{private}$ threshold:

$$D_{private} \geq S \cdot \frac{I + (1 - \pi)D_{social}}{1 - S}$$

If actors are forced to post collateral or are exposed to private financial liability that satisfies this condition, then the suicide region closes and deployment of AGI will occur only when NPV is positive, making the rationale for delaying deployment worthwhile.

5.2. Relaxing the Winner-Takes-All Conditions (Windfall Clauses)

Rather than targeting the private liability of players ($D_{private}$), we can instead focus on adjusting the market share parameter (S) to arrive at a more socially optimal solution. In our baseline model, the strict winner-takes-all condition ($S \approx 0$) drives the denominator of V_P^* towards its maximum value: π . Subsequently, this lowers the preemption threshold (V_P^*) and induces early deployment of AGI.

However, if the prize of successful (aligned) AGI deployment is shared, V_P^* can increase to safer levels. For example, if the spoils of successful AGI adoption are shared equally among rivals ($S = 0.5$), then $V_P^* \rightarrow \infty$ and the suicide region ceases to exist. When $V_P^* > V_S^*$, the decision to deploy AGI will be based on NPV analysis that incorporates the probability of misalignment, encouraging the implementation of optimal safety testing and protocols.

A “windfall clause”, as suggested by O’Keefe et al. (2020), may provide such a solution where S is non-negligible. Essentially, it is a treaty mechanism that redistributes economic rents of AGI ex-post, effectively eliminating the incentive for early deployment. By guaranteeing the second mover with a non-zero-payoff, the “fear of missing out” is dampened while allowing the “fear of error” to dominate once again, increasing the value of the option to wait.

5.3. The Role of Verifiability (Reducing Strategic Uncertainty)

A primary driver of the urgency to deploy early is the uncertainty of a rival’s development status: an actor is unsure of how close their rival is to deploying AGI. In our model, strategic uncertainty compels a player to race towards early deployment and erodes the present value of the option to wait. However, if there exists a robust verification process (e.g. hardware-level verification or blockchain-based monitoring), this will alter the game from being one of imperfect information with high strategic uncertainty to one of perfect information. With perfect monitoring of rivals, two modifications to player behaviours are predicted:

A. Restoration of the Monopolist Interval (Stability)

If monitoring is reliable and confirms that a rival will remain technically incapable of deploying AGI for a lengthy period, this completely disarms the immediate threat of preemption. During this interval, the Leader gains an effective monopoly which allows them to ignore the competitive pressure ($V_t > V_P^*$) and revert to a rational NPV calculus ($V_t > V_S^*$). As long as the Leader is confident in the verification process and that its rival is not technically able to deploy AGI, then – even though they may be situated in the suicide region – they will choose to “step off the ledge” and prioritize safety.

B. The "Breakout" Instability

However, perfect information introduces a sprint towards deployment during the endgame phase of the arms race. If a lagging agent is made aware of its rival being close to the threshold that makes AGI deployment viable ($V_t > V_S^*$) then – in a preemptive move to prevent their imminent total loss of hegemony - the slower agent may deploy a premature unsafe AGI system in self-defense.

While effective verification may eliminate the suicide region during the early stages of AGI research when neither player is close to being technically capable, it does not prevent suboptimal (unsafe) deployment of AGI as deployment capability increases. This is due to the inherent winner-takes-all dynamic ($S \approx 0$) that compels agents to launch pre-maturely when faced with imminent loss.

5.4. The Limitations of “Warning Shots”

A common prediction for the arms race is that a “warning shot” is likely to occur: a limited, sub-existential disaster that will serve as a “Chernobyl moment” for AGI¹⁶. It is posited that such an event will trigger global cooperation and lead to actors towards reprioritizing safety over expedience (Christiano, 2019). However, our model predicts that such a warning shot is unlikely to have a decisive impact on the dynamics of the race, unless it results in a change to the underling payoff structures of its players. This is because the cost of disaster (D) appears symmetrically in the Leader-Follower payoffs, and so that, in equilibrium, the indifference condition ($L = F$) cancels the impact of D on both sides of equation (the “cancellation effect”). That is, an increase in the magnitude of D does not materially impact the decision to preemptively deploy AGI.

If a warning shot occurs, it may cause players to revise their posterior belief of D upwards, but due to the winner-takes-all nature of the race ($S \approx 0$), little will change in terms of the pressure

¹⁶ Potential warning shots include: (1) an algorithmic flash crash occurs when an autonomous trading agent, instructed simply to maximize portfolio returns, triggers a global liquidity cascade to manipulate prices – demonstrating how a benign objective function can lead to systemic financial collapse via instrumental convergence; (2) a critical infrastructure failure, such as an optimization model cutting power to hospitals to balance grid load; and (3) a bio-security breach, where non-state actors utilize open-source models to synthesize novel pathogens. While these events would empirically confirm the non-zero probability of ruin ($D > 0$), the model suggests they would not alter the preemption equilibrium outcome.

for preemptive deployment. In fact, if agents believe that $\pi_{self} > \pi_{rival}$, a warning shot will accelerate the race (see equation (7)). A warning shot will only be effective if it is severe enough to trigger players and regulators to install the external liability mechanisms discussed in Section 5.1 and/or windfall clauses. In conclusion, a warning shot – in isolation – will not change the fundamental dynamics of the race.

6. Conclusion

This paper extends the theory of real options to account for endogenous existential risk. We show that when the prize is privatized but the cost of ruin is shared globally, the standard adage that risk induces caution no longer applies. We derive a region – coined the “suicide region” – where actors in the game will race towards early deployment of a high-risk technology with a non-negligible probability of existential catastrophe despite a negative net present valuation. At face value, this result appears irrational. However, we show that the shared cost of global ruin results in a “cancellation effect” across the payoffs of the first and second mover such that the expected loss of systemic ruin plays no role in the calculus of exercising the option to deploy. The current geopolitical trajectory in the AGI race is not an anomaly, but rather a Nash equilibrium in a pathological game.

We conclude that voluntary coordination solutions are insufficient in resolving the suicide region puzzle. It is the underlying fundamental payoffs that must be altered. We suggest mechanisms that result in non-trivial private liability of AGI misalignment (systemic ruin) – such as catastrophe bonds, escrow accounts, strict tort liability – and shared benefits of AGI success – such as windfall clauses – to restructure the payoffs of the players. By revising the payoffs with such mechanisms, the option to delay AGI deployment increases, reducing the aggregate risk of catastrophe. Effective monitoring may also delay deployment, though it is predicted to force players to race at the end game if the winner-takes-all nature of the payoffs remain intact.

Ultimately, our findings map the AGI race onto the political economy framework of "The Narrow Corridor" (Acemoglu & Robinson, 2019). The path to safe deployment of AGI requires a

“Red Queen” dynamic¹⁷ where the advancement in AI safety research (society’s shackles) must scale proportionally with technical capabilities. Currently, the incentive structure of the game ($L > F$) compels players towards early preemptive deployment of AGI without the corresponding safety mechanism in place ($V_P^* < V_t < V_S^*$), pushing the world out of the “narrow corridor” and toward the despotic outcome of an unshackled Leviathan – an AGI system unmoored from human control and prone to systemic misalignment. Re-entering the corridor requires not only technological progress, but mechanisms that are designed to require safety research as prerequisites for economic viability.

REFERENCES

- Acemoglu, D., & Robinson, J. A. (2019). *The Narrow Corridor: States, Societies, and the Fate of Liberty*. New York, NY: Penguin Books.
- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.
- Armstrong, S., Bostrom, N., & Shulman, C. (2016). Racing to the precipice: A model of artificial intelligence development. *AI & Society*, 31(2), 201–206.
- Aschenbrenner, L. (2024). *Situational Awareness: The Decade Ahead*. Retrieved from <https://situational-awareness.ai/>
- Bartlett, L. (Host). (2023, July). *Dario Amodei, CEO of Anthropic* [Audio podcast episode]. In *The Logan Bartlett Show*.
- Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford, UK: Oxford University Press.
- Bostrom, N., Douglas, T., & Sandberg, A. (2016). The unilateralist’s curse and the case for a principle of conformity. *Social Epistemology*, 30(4), 350–371.
- Christiano, P. (2019). What failure looks like. AI Alignment Forum. Retrieved from: <https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like>
- Dixit, A. K., & Pindyck, R. S. (1994). *Investment under Uncertainty*. Princeton, NJ: Princeton University Press.
- Good, I. J. (1965). Speculations concerning the first ultraintelligent machine. *Advances in Computers*, 6, 31–88.

¹⁷ The term refers to the 'Red Queen' effect originally proposed in evolutionary biology by Van Valen (1973), derived from Lewis Carroll’s *Through the Looking-Glass*: “Now, here, you see, it takes all the running you can do, to keep in the same place”. Acemoglu and Robinson (2019) adapt this to political economy, arguing that liberty is not a static condition but a dynamic struggle where civil society’s constraints must evolve at the same velocity as the state’s coercive power.

- Goertzel, B. (2014). Artificial general intelligence: Concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1), 1–46.
- Grace, K., Stewart, H., Sandkühler, J. F., Thomas, S., Weinstein-Raun, B., & Brauner, J. (2024). Thousands of AI authors on the future of AI. *arXiv preprint arXiv:2401.02843*.
- Grenadier, S. R. (2002). Option exercise games: An application to the equilibrium investment strategies of firms. *The Review of Financial Studies*, 15(3), 691–721.
- Huisman, K. J., & Kort, P. M. (2003). Strategic investment in technological innovations. *European Journal of Operational Research*, 144(1), 209–223.
- Jones, C. I. (2024). The A.I. dilemma: Growth vs. existential risk. *American Economic Review Insights*, 6(4), 575-590.
- Lubetkin, D. (2024, May 17). OpenAI dissolves team focused on long-term AI risks. *The Wall Street Journal*.
- Metz, C. (2023, May 1). 'The Godfather of A.I.' Leaves Google and Warns of Danger Ahead. *The New York Times*. Retrieved from <https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html>
- Naudé, W., & Gries, T. (2022). Breakthroughs, Backlashes and Artificial General Intelligence: An Extended Real Options Approach. *IZA Institute of Labor Economics*, 15077.
- Nordhaus, W. D. (2021). Are we approaching an economic singularity? Information technology and the future of economic growth. *American Economic Journal: Macroeconomics*, 13(1), 299–332.
- O'Keefe, C., Cihon, P., Garfinkel, B., Flynn, C., Leung, J., & Dafoe, A. (2020). The windfall clause: Distributing the benefits of AI for the common good. *Centre for the Governance of AI*.
- Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. New York, NY: Viking.
- Suleyman, M. (2023). *The Coming Wave: Technology, Power, and the Twenty-first Century's Greatest Dilemma*. New York, NY: Crown.
- The Economist. (2024a, April 25). China and America are locked in a race to become the world's leading AI power. *The Economist*.
- The Economist. (2024b, May 26). OpenAI's former board members: The company cannot be trusted to govern itself. *The Economist*.
- Van Valen, L. (1973). A new evolutionary law. *Evolutionary Theory*, 1, 1–30.
- Weeds, H. (2002). Strategic delay in a real options model of R&D competition. *The Review of Economic Studies*, 69(3), 729–747.