Racing Against Tomorrow's Safeguards*

Pavel Kocourek†

May 18, 2026

Abstract

I study whether an AI race is necessarily an arms race, using a Loury (1979)-style patent-race model in which the probability of catastrophe conditional on invention falls over time as society becomes more prepared. Even with effort costs set to zero, firms voluntarily slow down: racing faster brings invention forward into a higher-risk window, so a lab acting purely in its own interest holds back. Under initial certainty, a unique interior Nash equilibrium emerges in closed form and dissipates the entire prize through catastrophe risk. When initial risk is strictly below one, a Pareto-superior slowdown equilibrium coexists with a corner racing equilibrium, so the binding policy task becomes coordination rather than unilateral restraint — and "arms race" rhetoric pushes labs toward the inferior outcome.

1 Introduction

AI development is widely described as a race, but is it an arms race? Hendrycks et al. (2023) organise catastrophic AI risks into four categories and name the "AI race" as one of them: a dynamic in which competitive environments compel actors to deploy unsafe AIs. The answer shapes the scope for voluntary restraint: in an arms race, building is a dominant strategy and no unilateral slowdown is individually rational; in a coordination game, mutual restraint can be self-enforcing. Armstrong et al. (2016) study a contest in which teams trade off capability against safety and show that competitive pressure can drive equilibrium safety investment to dangerously low levels. Grace (2022) organizes the strategic possibilities into a compact taxonomy of three toy games — an *arms race* (building is dominant), a *safety-or-suicide race* (a coordination game in which labs build only if the other does), and a *suicide race* (where the damage from building is so severe that mutual standstill prevails) — that hinge on the probability of catastrophe conditional on building.

All three of Grace's games treat building as binary: build or don't.[1] But AI development is not on/off — labs also choose *how fast* to push.

*Preliminary and incomplete — please do not circulate.

†Mercator School of Management, University of Duisburg–Essen. Email: pavel.kocourek@uni-due.de.
I am grateful to Katja Grace for mentorship and guidance, to Yiman Sun for helpful comments on an early discussion of the idea, and to the SPAR (Supervised Program for Alignment Research) fellowship for support. All errors are my own.

[1]The model's two symmetric players can equally be read as states competing for AI supremacy, though the symmetry and shared-catastrophe assumptions are more defensible at the lab level.

Speed matters because the world is not standing still. Alignment research advances, interpretability tools improve, evaluations get sharper, and governance catches up. In other words, society is progressively more prepared for advanced AI, so the probability of catastrophe conditional on invention falls over time.[2] Growing preparedness creates an incentive to wait: racing faster brings invention forward into a higher-risk window, while waiting lets the same invention arrive into a safer one. A voluntary slowdown, by a lab acting purely in its own interest, becomes a real possibility.

This paper asks whether that incentive alone is strong enough to make labs voluntarily slow down:

> *Is growing preparedness, by itself, enough to discipline a competitive innovation race — even when effort costs are zero?*

In the canonical Loury (1979) patent race and its successors, strictly convex effort costs are what pin down an interior equilibrium. Remove the costs and the standard model collapses: firms simply race at the maximum feasible speed. I embed catastrophic risk as a symmetric negative externality in this costless environment and study whether a time-declining catastrophe probability can do the disciplining work that convex costs normally do. Zero cost is not a realism claim — R&D is obviously expensive — but a deliberate stress test: I strip costs out to isolate the catastrophic-risk channel. If the slowdown result survives under entirely free effort, it survives *a fortiori* under positive costs.

The analysis yields two results, presented in increasing order of generality: a sharp characterization under initial certainty, then a relaxation that drops it.

Under a time-decreasing catastrophe probability $p(t) = e^{-\delta t}$, so that catastrophe is initially certain, a genuinely interior symmetric Nash equilibrium emerges, with closed-form combined effort $\lambda^* = x_1^* + x_2^* = \frac{1}{2}V\delta/D$ — each lab choosing $x_i^* = V\delta/(4D)$, where $V > 0$ is the prize, $D > 0$ is the cost of catastrophe, and $\delta > 0$ is the rate at which catastrophe risk decays (Proposition 1) — and zero expected profits in the spirit of Loury (1979) (Corollary 1). Firms voluntarily choose a finite, interior speed despite effort being entirely free; the disciplining force is catastrophic risk, not cost.

Relaxing initial certainty to $p_0 \in (0, 1)$, the slowdown equilibrium does not vanish but is joined by a corner racing equilibrium (Proposition 2). Whenever both exist, the slowdown profile strictly Pareto-dominates racing (Corollary 3), so the policy question shifts from whether unilateral restraint is rational to whether labs coordinate on the better equilibrium — a coordination problem that "arms race" rhetoric actively pushes in the wrong direction. This is the result with the most direct bearing on current debate: public discussion routinely treats mutual restraint as something that would require intense international cooperation beyond anything observed, but the model shows mutual restraint can be self-enforcing once one side is seen to slow.

A side observation (Corollary 2) is that the equilibrium catastrophe probability in the unique-equilibrium case is independent of δ: anticipated improvements in societal pre-

[2] In reality the capability target itself moves: each generation of advanced AI is more capable than the last, so preparedness for the systems labs are actually about to build need not improve over calendar time and could even deteriorate. The model abstracts from this by holding the capability target fixed and asking whether, in that simpler world, a slowdown incentive arises at all. The fixed-target reading is closer to "the next major capability jump" than to any timeless notion of AGI.

paredness are partly absorbed by more intense racing rather than translated into safer outcomes. The result is sharp under the model's no-cost, full-information specification and weakens once those assumptions are relaxed, so it is best read as a qualitative warning rather than a quantitative claim.

1.1 Related literature

The model sits at the intersection of two largely disconnected literatures. I summarise the closest prior work here.

Patent race theory. Loury (1979) is the canonical stochastic race with Poisson innovation and convex effort costs; Lee and Wilde (1980) and Reinganum (1982) develop flow-cost and continuous-time variants, and Reinganum (1989) surveys the field. All of these papers assume strictly convex costs and feature no catastrophic externality. To the best of my knowledge, the zero-cost case has never been studied because, without an externality, it is trivial. A related strand is Schumpeterian growth theory: Aghion and Howitt (1992) use the same Poisson innovation technology with creative destruction as a negative externality that disciplines equilibrium R&D. The present paper substitutes catastrophic risk for creative destruction, but the disciplining logic is analogous.

AI race models. Armstrong et al. (2016) study a one-shot game in which teams choose safety levels, racing for an AI prize. Their model studies how competition erodes safety investment in a one-shot setting. Grace (2022) introduces a distinct three-regime taxonomy of racing incentives — an arms race, a safety-or-suicide race, and a suicide race — that is static and has no continuous effort choice. Hendrycks et al. (2023) provide an informal-taxonomy counterpart that names the "AI race" as one of four categories of catastrophic AI risk; the present paper supplies the game-theoretic foundation for that category. Tan (2025), "The Suicide Region," is the closest continuous-time paper: he models AGI deployment as an option-game preemption race and identifies a region in which competitive pressure forces rational deployment despite negative risk-adjusted net present value. Tan's framework and mine are complementary. Tan studies binary deployment timing with symmetric catastrophic risk; his key result is that the shared ruin cost *cancels* from the equilibrium indifference condition, so catastrophe severity does not deter the race. I study continuous effort with time-decaying catastrophic risk; my key result is the opposite — catastrophe *constrains* equilibrium effort and pins down a finite interior solution. The two results are not in tension: they come from different modelling choices, and together they sketch the boundary of when catastrophic risk slows the race and when it does not.

Safety and growth. Jensen et al. (2023) study a static contest where agents allocate between safety and performance spending; their margin is different from mine (safety-vs-performance rather than speed). Jones (2024) derives the growth-vs-risk tradeoff for a single planner and characterizes when optimal AI progress should continue or halt; there is no strategic interaction. My paper complements Jones by asking what happens when his tradeoff is decentralized to two competing firms.

2 Model Setup

Environment. Two symmetric firms $i \in \{1, 2\}$ simultaneously and irrevocably choose effort levels $x_i \in [0, \bar{x}]$ at time $t = 0$. Effort is costless ($c(x) \equiv 0$), the discount rate is zero, and the cost of catastrophe is borne symmetrically by both firms. These simplifying assumptions strip away the standard cost-based and discounting channels so as to isolate the pure slowdown incentive generated by catastrophic risk.[3]

Innovation technology. Each firm's innovation arrives as an independent Poisson process with rate equal to its effort: $T_i \sim \text{Exp}(x_i)$. The arrival time of the first innovation is $T = \min(T_1, T_2) \sim \text{Exp}(\lambda)$, where $\lambda \equiv x_1 + x_2$ is the aggregate effort. Conditional on positive efforts, firm i wins the race with probability x_i/λ; under symmetric play this reduces to $\frac{1}{2}$.

Catastrophic risk. Catastrophe is a single event tied to the *first* arrival in the race: when the first innovation occurs at time $T = \min(T_1, T_2)$, catastrophe materialises with probability $p(T)$ and otherwise does not. There is no separate, independent catastrophe draw for the second arrival or for each player — $p(t)$ should be read as the (Bernoulli) probability that the world's hidden "advanced AI is dangerous" state turns into an actual catastrophe given that advanced AI has just arrived at calendar time t.[4] The catastrophe probability declines as society grows more prepared:

$$p(t) \;=\; p_0\, e^{-\delta t}, \qquad p_0 \in (0, 1], \quad \delta > 0,$$

with initial catastrophe probability $p_0 \equiv p(0)$ and preparedness improving at rate δ. Section 3 treats the sharp case of initial certainty, $p_0 = 1$; Section 4 relaxes it to $p_0 \in (0, 1)$. A constant catastrophe probability — preparedness that never improves — is the degenerate $\delta \to 0$ boundary of this family; Section 3 opens by showing it supports no interior equilibrium under costless effort, which is precisely why the time-declining force is what does the disciplining work.

Payoffs. Conditional on innovation at time t:

Event	Probability	Winner	Loser
No catastrophe	$1 - p(t)$	V	0
Catastrophe	$p(t)$	$-D$	$-D$

The parameter $V > 0$ is the prize from being first and $D > 0$ is the cost of catastrophe; because D is borne by both firms whenever catastrophe strikes, catastrophe enters the model as a *symmetric* negative externality.[5]

[3] The upper bound $\bar{x}$ is a compactness device: without it, the win probability x_i/λ is ill-defined as $\lambda \to \infty$ under symmetric play. I take $\bar{x}$ large enough that it does not bind at the interior equilibrium.

[4] I use "advanced AI" rather than "AGI" to denote a fixed capability target whose arrival triggers both the prize V and the catastrophe risk $p(T)$; the analysis is agnostic about which threshold, provided it is held fixed.

[5] A non-zero loser no-catastrophe payoff L maps into the present model by a constant utility shift: the game reduces to the present one with $(V, D) \mapsto (V - L,\, D + L)$, valid for $L \in (-D, V)$. Positive L (knowledge spillovers) shrinks the effective V/D ratio and expands the slowdown region of Section 4; negative L (winner-take-all dynamics) contracts it. This case is therefore already covered by the existing analysis.

3 Slowdown Without Costs

Why timing has to vary. Start from the degenerate case of a constant catastrophe probability, $p(t) = p$. Since p does not depend on the arrival time, firm i's expected payoff factors into

$$U_i(x_i, x_j) \;=\; \frac{x_i}{\lambda}(1-p)V \;-\; pD, \qquad \frac{\partial U_i}{\partial x_i} \;=\; \frac{x_j}{\lambda^2}(1-p)V \;>\; 0$$

whenever the rival exerts any effort. Aggregate effort λ then affects only *when* the race ends, not *what* the players get in expectation, so with costless effort each firm races to the cap and no interior equilibrium exists. A time-declining catastrophe probability is exactly what breaks this degeneracy: it makes the expected payoff depend on *when* invention arrives, and hence on how fast the labs race.

Accordingly, let $p(t) = e^{-\delta t}$, so that $p_0 = 1$ and societal preparedness improves at rate δ. Time now re-enters the payoff through the arrival distribution. Throughout this section I assume $\bar{x} > V\delta/(4D)$, so that the interior equilibrium lies strictly below the upper bound.

3.1 Expected payoff

Innovation arrives at $T \sim \text{Exp}(\lambda)$, so the probability of catastrophe at the first arrival is

$$\phi(\lambda) \;\equiv\; \mathbb{E}[p(T)] \;=\; \int_0^\infty e^{-\delta t}\,\lambda e^{-\lambda t}\,dt \;=\; \frac{\lambda}{\lambda+\delta}.$$

This is strictly increasing and concave in λ, with $\phi(0) = 0$ and $\phi(\lambda) \to 1$ as $\lambda \to \infty$. The key observation is that *aggregate racing speed and catastrophe risk are monotonically linked*: a faster race brings arrival forward into a higher-risk window. Substituting into firm i's payoff (derivation omitted; see (Loury, 1979) for the standard steps),

$$U_i(x_i, x_j) \;=\; \frac{x_i}{\lambda}\cdot\frac{V\delta}{\lambda+\delta} \;-\; \frac{\lambda D}{\lambda+\delta} \;=\; \frac{1}{\lambda+\delta}\left[\frac{x_i}{\lambda}\,V\delta \;-\; \lambda D\right], \qquad \lambda = x_i + x_j.$$

The interpretation is transparent. With probability x_i/λ, firm i wins the race and receives the prize V provided catastrophe does not strike (probability $\delta/(\lambda+\delta)$). Regardless of who wins, catastrophe occurs with probability $\lambda/(\lambda+\delta)$ and imposes cost D on both firms. Because both probabilities depend on λ, time-variation in catastrophe risk has not dropped out — this is exactly the force the constant-p case lacked.

3.2 Equilibrium

I look for symmetric Nash equilibria $x_1 = x_2 = x^*$, with aggregate effort $\lambda^* = 2x^*$. Differentiating U_i with respect to x_i at fixed x_j, and using

$$\frac{\partial}{\partial x_i}\left(\frac{x_i}{\lambda}\right) = \frac{x_j}{\lambda^2}, \qquad \phi'(\lambda) = \frac{\delta}{(\lambda+\delta)^2},$$

the first-order condition is

$$\frac{\partial U_i}{\partial x_i} \;=\; \frac{x_j}{\lambda^2}\,V\big[1-\phi(\lambda)\big] \;-\; \frac{x_i}{\lambda}\,V\,\phi'(\lambda) \;-\; \phi'(\lambda)\,D \;=\; 0.$$

Proposition 1 (Slowdown without costs)**.** *The costless patent race with decreasing catastrophe probability $p(t) = e^{-\delta t}$ has a unique Nash equilibrium. It is symmetric and interior, with*

$$x_i^* = \frac{V\delta}{4D}, \qquad \lambda^* = \frac{V\delta}{2D}.$$

Each lab chooses individual speed $x_i^* = V\delta/(4D)$; the *combined* racing speed of the two labs is $\lambda^* = x_1^* + x_2^* = \frac{1}{2}V\delta/D$, the form quoted in the introduction. At the interior equilibrium, each firm's marginal competitive gain from raising x_i — winning with higher probability — exactly offsets the marginal increase in expected catastrophe cost from pushing arrival into a higher-risk window. Symmetry of the first-order conditions then forces $x_1 = x_2$, and a standard concavity argument gives uniqueness. The formal proof is in Appendix A.

Substituting $\lambda^* = \frac{1}{2}V\delta/D$ back into the symmetric expected payoff yields a striking rent-dissipation result.

Corollary 1 (Zero expected profit)**.** *Each firm earns zero expected payoff: $U_i^* = 0$.*

The prospect of the prize V is competed away in full by the catastrophic-risk externality: in Loury, firms burn resources to compete; here, they burn the world. The rent-dissipation mechanism has shifted from private waste to public catastrophe.

3.3 An observation on δ and welfare

Substituting λ^* into $\phi(\lambda) = \lambda/(\lambda + \delta)$ pins down the equilibrium catastrophe probability.

Corollary 2 (Equilibrium probability of catastrophe)**.** *The equilibrium probability of catastrophe is*

$$\phi(\lambda^*) = \frac{V}{V + 2D},$$

independent of the preparedness rate δ.

This is a qualitative warning, not a quantitative prediction. The exact δ-independence holds only under the model's deliberately strong no-cost, full-information, full-commitment assumptions; it weakens as soon as any of them is relaxed. Read literally as "safety progress never lowers risk" it would be false — the claim is the softer one that part of anticipated preparedness is offset by faster racing. With that caveat fixed: the equilibrium catastrophe probability depends only on the ratio V/D, not on δ, because faster preparedness raises the equilibrium racing speed λ^* in exact proportion, leaving catastrophe risk untouched.
Assumptions behind the observation. The independence is sharp because the model imposes (i) zero effort costs, so the only check on speed is the catastrophe externality itself; (ii) full information about δ at $t = 0$, so labs internalise the entire trajectory of $p(t)$ when committing to effort; (iii) commitment, since effort is chosen once and cannot be revised when new safeguards arrive; and (iv) a single common δ that both labs respond to symmetrically. Each of these can be relaxed and the offset between δ and equilibrium speed will weaken. With positive R&D costs, in particular, equilibrium effort no longer scales linearly in δ, so faster preparedness produces some net reduction in catastrophe probability rather than full absorption. So, consistent with the caveat above: insofar as δ is anticipated by competing labs, part of its effect is absorbed by faster racing rather than translated into safer outcomes.

Welfare comparison. The social planner who internalizes both firms' catastrophe cost chooses $\lambda \geq 0$ to maximize joint expected profit

$$W(\lambda) \;=\; V\Big[1-\phi(\lambda)\Big] \;-\; 2D\,\phi(\lambda) \;=\; V \;-\; (V+2D)\,\phi(\lambda).$$

Since W is strictly decreasing in λ, the first-best at zero discount rate is the corner $\lambda^{\mathrm{SP}} \to 0^{+}$: the planner innovates at a vanishingly small rate and captures nearly the full prize V.[6] At the Nash equilibrium, by Corollary 1, $W(\lambda^{*}) = 2U_i^{*} = 0$: competition dissipates the full value of the prize through the catastrophe externality — a 100% welfare loss relative to the first-best $W \to V$.

4 Multiple Equilibria Without Initial Certainty

The preceding sections assume catastrophe is initially certain, $p_0 = 1$. This section relaxes that to $p_0 \in (0,1)$ and shows that the slowdown logic of Section 3 now produces a coordination problem with two symmetric Nash equilibria rather than a unique interior outcome. Throughout, $p(t) = p_0 e^{-\delta t}$ with $\delta > 0$, and $\bar{x}$ is taken to be sufficiently large but finite. Then

$$\phi(\lambda) \;=\; \frac{p_0\,\lambda}{\lambda+\delta}, \qquad \phi'(\lambda) \;=\; \frac{p_0\,\delta}{(\lambda+\delta)^2}, \qquad U_i(x_i,x_j) \;=\; \frac{x_i}{\lambda}\,V\Big[1-\phi(\lambda)\Big] \;-\; \phi(\lambda)\,D,$$

with $\lambda = x_i + x_j$. Evaluating $\partial U_i/\partial x_i$ at the symmetric profile $x_i = x_j = x$ ($\lambda = 2x$) and rescaling by the positive factor $4x(2x+\delta)^2$ gives the quadratic

$$P(x) \;\equiv\; 4V(1-p_0)\,x^2 \;+\; 4\delta\Big[V(1-p_0) - p_0 D\Big]\,x \;+\; V\delta^2,$$

whose sign equals the sign of the symmetric marginal payoff in own effort at (x,x).

Proposition 2 (Two equilibria with initial risk below one)**.** *Suppose $p_0 \in (0,1)$ and $\bar{x}$ is sufficiently large (specifically, $\bar{x} > x_-$, with x_- given in the proof). Let $r \equiv V/D$. Then:*

1. *The racing profile $(\bar{x},\bar{x})$ is a symmetric Nash equilibrium.*
2. *If in addition*

$$p_0 \;>\; \underline{p}(r) \;\equiv\; 1 \;-\; \frac{1}{(1+r)^2}, \tag{1}$$

then there is also a symmetric interior Nash equilibrium (x_-,x_-) with $x_- < \bar{x}$, in which both firms slow down.

Corollary 3 (Pareto dominance of the slowdown equilibrium)**.** *Whenever the slowdown equilibrium (x_-,x_-) of Proposition 2 exists, it strictly Pareto-dominates the racing equilibrium $(\bar{x},\bar{x})$.*

The proof is in Appendix A.

Equilibrium selection therefore becomes a substantive policy task: communication, reputation, and industry agreements coordinate beliefs on the Pareto-superior slowdown profile rather than force firms off an individually irrational outcome. The threshold $\underline{p}(r)$ is independent of δ, extending the preparedness treadmill of Corollary 2 to the coordination setting.

[6] This corner solution reflects the assumption of a zero discount rate.

Mapping onto Grace's taxonomy. Grace (2022) casts the AI race as one of three static 2×2 games — an *arms race* (building dominant), a *safety-or-suicide race* (a coordination game in which each lab builds only if the other does), and a *suicide race* (mutual standstill dominant). Proposition 2 reproduces the first two not as assumed payoff matrices but as endogenous outcomes of a dynamic, continuous-effort race: below the threshold $\underline{p}(r)$ only the racing equilibrium survives — Grace's arms race — while above it the slowdown and racing equilibria coexist with the slowdown strictly Pareto-dominant — exactly her safety-or-suicide coordination game. The contribution is not to supply game theory where Grace had none; she already works with a payoff matrix. It is to put her taxonomy on a *dynamic* footing, in which the prevailing regime is *derived* from the prize-to-damage ratio r and the initial risk p_0, and the slowdown force is generated by growing preparedness rather than assumed. Grace's strict suicide race does not arise here: with zero effort cost a sole entrant can always capture the catastrophe-adjusted prize, so mutual standstill is never strictly dominant. Recovering it requires positive effort costs — a generalization deferred to future work (see "Planned extensions" below).

Proof. Part (1). The leading coefficient of P is $4V(1-p_0) > 0$, so $P(\bar{x}) \to +\infty$ as $\bar{x} \to \infty$; for $\bar{x}$ large, $P(\bar{x}) > 0$ and the symmetric marginal at $(\bar{x}, \bar{x})$ is strictly positive. Each firm's unconstrained best response would exceed $\bar{x}$; constrained by the cap, the best response is $\bar{x}$ itself, so $(\bar{x}, \bar{x})$ is Nash.

Part (2). Since P opens upward with $P(0) = V\delta^2 > 0$, the quadratic dips below zero on some interval iff its discriminant is strictly positive. A direct computation reduces the discriminant condition to

$$p_0 D^2 \;>\; V(1-p_0)(V+2D).$$

Dividing by $D^2(1-p_0) > 0$ and writing $r = V/D$ yields $p_0/(1-p_0) > r(r+2)$, which rearranges to $p_0 > r(r+2)/(1+r)^2 = 1 - 1/(1+r)^2 = \underline{p}(r)$ (using $(1+r)^2 = 1 + r(r+2)$), i.e., condition (1). Under (1), P has two distinct positive roots $x_- < x_+$, with $P > 0$ on $(0, x_-)$, $P < 0$ on (x_-, x_+), and $P > 0$ on (x_+, ∞). The symmetric marginal therefore flips from positive to negative as x crosses x_-, so along the symmetric diagonal the payoff of each firm attains a local maximum in own effort at x_-. It remains to verify unilateral best response: with $x_j = x_-$ held fixed, the first-order condition in x_i, after clearing the positive denominator, reduces to a polynomial of degree at most two in x_i, and the argument of Proposition 1 (strict negativity of the second derivative at any interior critical point, hence at most one such point) applies unchanged. Since $x_i = x_-$ satisfies the first-order condition by construction, it is the unique interior best response to x_-. Hence (x_-, x_-) is a symmetric Nash equilibrium. For $\bar{x}$ large enough, $x_- < \bar{x}$ and the profile is interior. □

Planned extensions. Two generalisations of Proposition 2 are deferred to a future revision: (i) introducing positive R&D costs, and (ii) relaxing the exponential form of $p(t)$ to a broader class of monotonically decreasing catastrophe paths. The qualitative story — two equilibria with one strictly slower whenever initial risk is high enough — is expected to survive both, and isolating which assumptions actually drive it is what the more general analysis is meant to deliver.

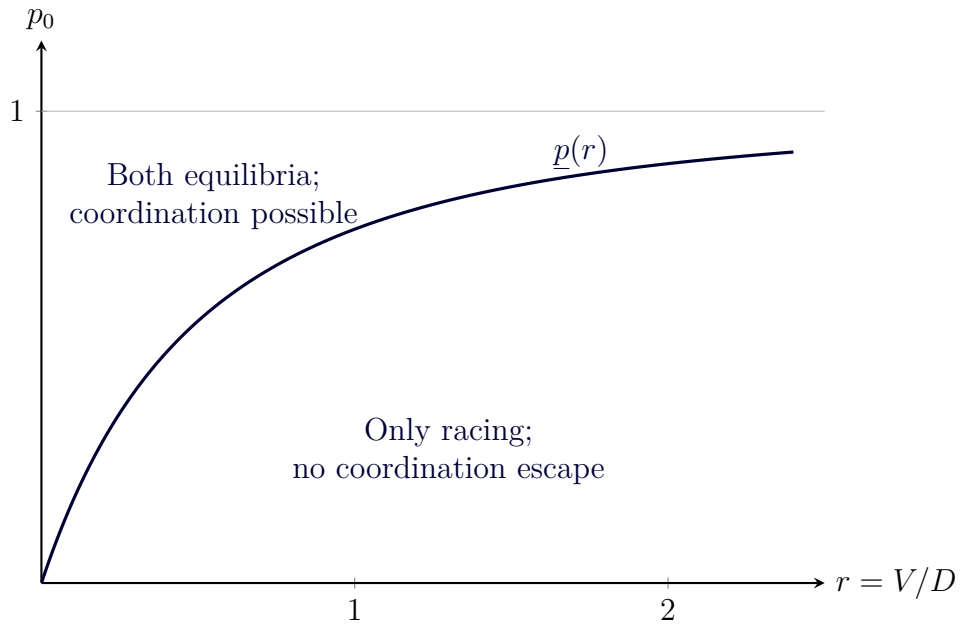

Figure 1: Region in the $(V/D,\, p_0)$-plane where the interior slowdown equilibrium of Proposition 2 exists. The boundary is $\underline{p}(r) = 1 - 1/(1+r)^2$ with $r = V/D$. Above the curve, both the slowdown equilibrium $(\underline{x}_-, \underline{x}_-)$ and the racing equilibrium $(\bar{x}, \bar{x})$ are Nash; below it, only the racing equilibrium exists.

5 Concluding Remarks

Catastrophic risk, on its own, can discipline a competitive innovation race: even with entirely free effort, firms voluntarily choose a finite interior speed because racing faster shifts invention into a higher-risk window. Under initial certainty this pins down a unique equilibrium with closed-form aggregate effort $\lambda^* = \frac{1}{2}V\delta/D$ and zero expected profits. Under initial risk strictly below one, a Pareto-superior slowdown equilibrium coexists with corner racing, so the binding question is no longer whether unilateral restraint is rational but whether labs coordinate on the better equilibrium. "Arms race" rhetoric works against that coordination: it pushes beliefs toward the inferior outcome that both labs would gladly leave.

A subsidiary observation is that, in the unique-equilibrium case, the equilibrium catastrophe probability $V/(V+2D)$ depends only on the ratio of prize to damage and not on the preparedness rate δ (Corollary 2). The result rests on the model's no-cost, full-information assumptions and weakens once those are relaxed; even so, it suggests that part of the safety improvement competing labs anticipate may be absorbed by faster racing rather than reflected in equilibrium risk. Two kinds of intervention shift the parameters that determine equilibrium risk in this model: those that move the payoffs — raising D via liability or lowering V via prize-dilution mechanisms such as mandatory licensing — and those that bind aggregate effort directly, such as compute caps or treaty-based moratoria. They operate alongside, rather than as substitutes for, work on alignment, interpretability, evaluations, and governance, which raise δ and thereby reduce risk along the path to whatever equilibrium effort the labs settle on.

A Proofs

Proof of Proposition 1

Proof. Let $(x_1, x_2) \in (0, \bar{x})^2$ be an interior Nash equilibrium and write $\lambda = x_1 + x_2$. Each firm's first-order condition holds with equality:

$$\frac{\partial U_i}{\partial x_i} \;=\; \frac{x_j}{\lambda^2}\, V\Big[1 - \phi(\lambda)\Big] \;-\; \frac{x_i}{\lambda}\, V\phi'(\lambda) \;-\; \phi'(\lambda)\, D \;=\; 0,$$

and firm j's FOC is the same with the indices swapped. Subtracting one from the other and factoring,

$$(x_j - x_i)\, \frac{V\Big[1 - \phi(\lambda) + \lambda\, \phi'(\lambda)\Big]}{\lambda^2} \;=\; 0.$$

Since $1 - \phi(\lambda) > 0$ and $\phi'(\lambda) > 0$ for every $\lambda > 0$, the numerator is strictly positive, so $x_1 = x_2 \equiv x^*$ and $\lambda^* = 2x^*$. Substituting back into the FOC,

$$\frac{V\Big[1 - \phi(\lambda^*)\Big]}{4x^*} \;=\; \phi'(\lambda^*)\Big(\frac{V}{2} + D\Big),$$

and using $\phi(\lambda) = \lambda/(\lambda + \delta)$ and $\phi'(\lambda) = \delta/(\lambda + \delta)^2$,

$$\frac{V\delta}{2\lambda^*(\lambda^* + \delta)} \;=\; \frac{\delta}{(\lambda^* + \delta)^2}\Big(\frac{V}{2} + D\Big).$$

Cancelling $\delta > 0$ and cross-multiplying by $2\lambda^*(\lambda^* + \delta)^2$ yields $V(\lambda^* + \delta) = V\lambda^* + 2D\lambda^*$, so $V\delta = 2D\lambda^*$ and hence $\lambda^* = \frac{1}{2}V\delta/D$. The per-firm effort $x^* = \lambda^*/2 = V\delta/(4D)$ lies in $(0, \bar{x})$ by the section-level assumption. Finally, uniqueness of each firm's best response follows from two observations. First, a direct calculation shows that at any interior critical point of $U_i(\cdot, x_j)$ with $x_j > 0$, the second derivative $\partial^2 U_i/\partial x_i^2$ is strictly negative, so every interior critical point is a strict local maximum. Second, clearing the positive denominator $\lambda^2(\lambda+\delta)^2$ in the FOC reduces it to a polynomial equation in x_i of degree at most two; two critical points would both be strict local maxima by the first observation, forcing a local minimum strictly between them and thus a third critical point, contradicting the quadratic bound. Hence $U_i(\cdot, x_j)$ has at most one interior critical point, which is the unique best response; the symmetric interior profile is therefore the unique Nash equilibrium of the game. □

Proof of Corollary 3

Proof. Let x_- be the smaller positive root of P from Proposition 2. The slowdown equilibrium payoff is

$$U_i^{\text{slow}} \;=\; \frac{V}{2} \;-\; \Big(\frac{V}{2} + D\Big)\, \phi(2x_-),$$

and the racing payoff as $\bar{x} \to \infty$ is $U_i^{\text{race}} = \frac{V(1-p_0)}{2} - p_0 D$. Since $\phi(2x_-) < p_0 = \lim_{\lambda\to\infty} \phi(\lambda)$, direct substitution gives $U_i^{\text{slow}} - U_i^{\text{race}} = \Big(\frac{V}{2} + D\Big)\Big(p_0 - \phi(2x_-)\Big) > 0$. □

References

Aghion, P. and Howitt, P. (1992). A model of growth through creative destruction. *Econometrica*, 60(2):323–351.

Armstrong, S., Bostrom, N., and Shulman, C. (2016). Racing to the precipice: A model of artificial intelligence development. *AI & Society*, 31(2):201–206.

Grace, K. (2022). Let's think about slowing down AI. AI Impacts blog post, 29 December 2022.

Hendrycks, D., Mazeika, M., and Woodside, T. (2023). An overview of catastrophic AI risks. *arXiv preprint*.

Jensen, M., Emery-Xu, N., and Trager, R. (2023). Industrial policy for advanced AI: Compute pricing and the safety tax. *arXiv preprint*.

Jones, C. I. (2024). The A.I. dilemma: Growth versus existential risk. *American Economic Review: Insights*, 6(4):575–590.

Lee, T. and Wilde, L. L. (1980). Market structure and innovation: A reformulation. *Quarterly Journal of Economics*, 94(2):429–436.

Loury, G. C. (1979). Market structure and innovation. *Quarterly Journal of Economics*, 93(3):395–410.

Reinganum, J. F. (1982). A dynamic game of R and D: Patent protection and competitive behavior. *Econometrica*, 50(3):671–688.

Reinganum, J. F. (1989). The timing of innovation: Research, development, and diffusion. In Schmalensee, R. and Willig, R. D., editors, *Handbook of Industrial Organization, Volume 1*, pages 849–908. North-Holland.

Tan, D. (2025). The suicide region: Option games and the race to artificial general intelligence. *arXiv preprint*.