Racing Against Tomorrow's Safeguards

Is the AI race really an arms race?

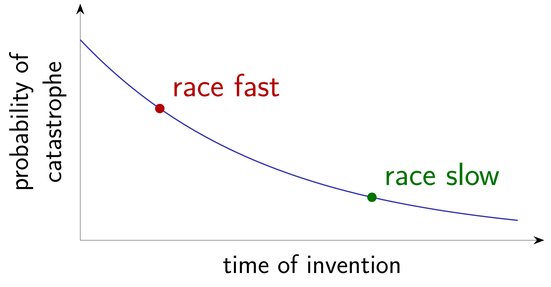
Pavel Kocourek | pkocourek.com | mentored by Katja Grace (AI Impacts)

Abstract

The AI race is widely called an **arms race**, a prisoner's dilemma: no lab can afford to hold back.

I question this framing. A catastrophe risk that **falls over time** disciplines the race on its own: racing faster pulls invention into the high-risk window, so even a self-interested lab holds back.

There are *two equilibria* — a reckless race and a calm slowdown everyone prefers — so the real problem is **coordination**.



Introduction

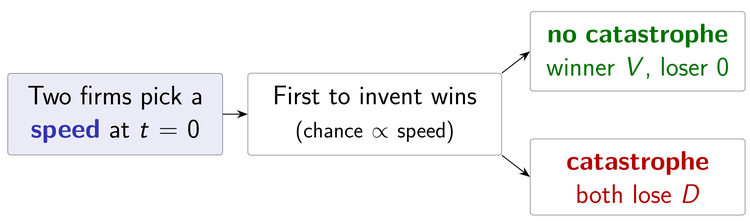
Katja Grace already argues, with a simple payoff matrix, that it is *not* an arms race (2022).

I put that on a dynamic footing: catastrophe risk falls as the world grows more prepared.

Racing faster pulls invention into the **high-risk window**; waiting lets it arrive safer.

The model

Two firms each pick a development **speed** at $t = 0$; the first to invent wins. Let λ be their *combined* speed.



A Loury (1979) patent race — with a chance of catastrophe added; effort is free.

Preparedness makes risk decay: $p_t = p_0\, e^{-\delta t}$.

Nash equilibrium: neither firm can do better by changing only its own speed.

Result 1 — the race slows itself

Risk is certain at the start ($p_0 = 1$). Even with *free* effort, firms hold back **voluntarily**, at $\lambda^* = \frac{1}{2} V\delta/D$.

The whole prize is competed away — not by effort, but by the catastrophe risk hurrying creates.

Preparedness treadmill: equilibrium catastrophe probability $V/(V + 2D)$, independent of δ — faster safety progress is absorbed by faster racing.

Result 2 — a calmer equilibrium

Catastrophe is likely, but *not* certain, at the start ($0 < p_0 < 1$) — now the game has **two symmetric equilibria**:

Everyone races	**Everyone slows**
a stable outcome — but reckless	also stable — and better for all

A calmer slowdown equilibrium exists — and it **Pareto-dominates** racing.

The coordination problem

Calling the AI race an arms race is a **self-fulfilling prophecy**: it coordinates beliefs on the bad equilibrium both labs would gladly leave.

A modest cap on development speed can tip firms onto the good equilibrium.

References

1. Glenn C. Loury (1979). Market structure and innovation. *Quarterly Journal of Economics*
2. Katja Grace (2022). *Let's think about slowing down AI*
3. Dan Hendrycks, Mantas Mazeika & Thomas Woodside (2023). *An Overview of Catastrophic AI Risks*
4. Stuart Armstrong, Nick Bostrom & Carl Shulman (2016). Racing to the precipice. *AI & Society*
5. McKay Jensen, Nicholas Emery-Xu & Robert Trager (2023). Industrial policy for advanced AI
6. Charles I. Jones (2024). The AI dilemma: growth versus existential risk
7. David Tan (2025). *The Suicide Region*