Racing Against Tomorrow's Safeguards

Is the AI race *really* an arms race?

Pavel Kocourek

pkocourek.com

SPAR Spring 2026 | mentor: Katja Grace (AI Impacts)

"Arms race" $\approx$ a prisoner's dilemma — each lab individually wants to race

Motivation

The AI race is widely seen as an **arms race**

a prisoner's dilemma: everyone races

"AI Is Not an Arms Race"

Katja Grace, TIME (2023) — argued informally

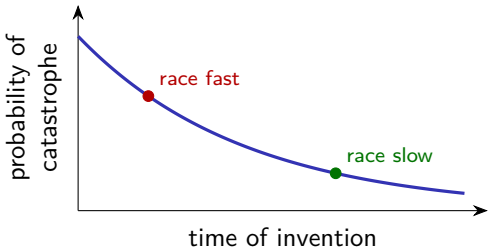
This paper puts it to a **dynamic** game-theoretic test

— and the dynamics only strengthen the case for slowing down

Preparedness changes the game

The world grows more prepared:

probability of catastrophe falls over time



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

A static race turns on the prize from winning vs. the catastrophe —
here, timing alone tips toward slowing

What's been done | what's new

Prior work

risk held fixed

build / don't-build

- Grace 2022
- Hendrycks et al. 2023

safety vs. speed

- Armstrong et al. 2016
- Jensen et al. 2023

growth vs. risk

- Jones 2024

dynamic timing, risk constant

- Tan 2025

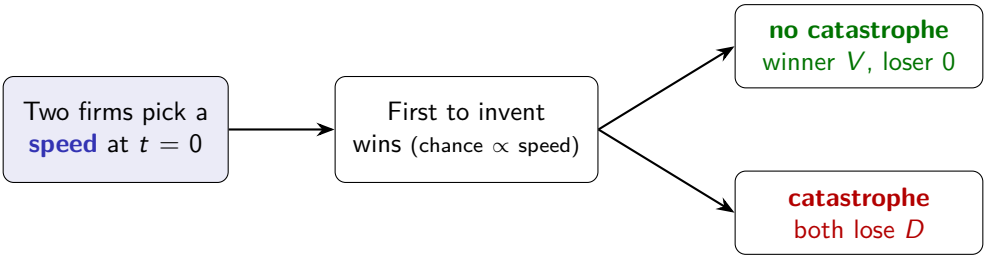
This paper

risk **falls over time**

The simplest dynamic race where catastrophe risk declines as the world grows prepared

$\rightarrow$ voluntary slowdown is an **equilibrium**, not unilateral restraint

The model



For a fixed target AI capability, the world keeps
getting more prepared — so catastrophe risk **falls over time**

Loury's (1979) patent race — with a chance of catastrophe added

Solution concept — Nash equilibrium

A pair of development speeds where

neither firm can do better

by changing only its own

Result 1 — the race slows itself

Assumption: catastrophe is **certain** at the start

They hold back **voluntarily**

— even if effort is free

- They compete every dollar away — **zero expected payoff**
- Faster preparedness only speeds up the race — catastrophe risk stays the same

Result 2 — a calmer equilibrium

Assumption: catastrophe is likely but **not certain** at the start

Now the game has **two equilibria**

Everyone races

stable: racing is each firm's best response

Everyone slows

stable too — and better for everyone

Either can take hold. The binding problem is **coordination**

Conclusion

Calling the AI race an arms race
is a **self-fulfilling prophecy**

The calmer equilibrium is real — the task is to
coordinate on it; even a modest speed limit can tip us there

pkocourek.com