THE FUTURE OF HUMANITY

Why Nobody Can Hit the Brakes

Researchers are leaving with warnings. Lab leaders put real odds on catastrophe. The race keeps going.

A public resignation, then others spoke

8 Sep

Jacob Coxon leaves Anthropic after earlier work at OpenAI

"Racing straight to self-improving superintelligence and gambling with our lives." Alignment leads at Anthropic added their own odds. Per Fortune and Time.

The chief executive states the odds

25%

"A chance that things go really, really badly"

Dario Amodei, Axios summit, September 2025. The CEO of a company that sells frontier models.

The numbers cannot be added up

More than 10%
Hubinger, within a decade

25%
Amodei, things go very badly

50%
Irving, extinction this decade

70%
Williams, without a slowdown

Different events, periods and definitions. Subjective estimates, as reported.

Slowing alone hands the lead to a rival

Both slow
Risk falls

One slows
The other leads

Both race
Risk is highest

Racing looks better for each lab whatever the other does. A prisoner's dilemma. Schematic.

Anthropic drops its pause commitment

24 Feb

2026, Responsible Scaling Policy v3.0

Reason given: catastrophic risk depends on all frontier developers, and slowing alone could let less careful actors lead.

A request for tools, not a pause

~1,300

lab employees signed "Pacing the Frontier", 28 July 2026

It asks Washington to back tools to pace automated AI research. Amodei followed on 12 September with a three-step plan. Trump rejected the idea.

METR finds the pace has doubled

~2x

Anthropic progress on well-specified tasks, October 2026

About 80% of it from AI itself, with substantial uncertainty. A majority of the key evidence was redacted. METR, 5 November.

You cannot end the race

Refusal line
What you will not deploy

Evidence
What vendors must show

Speed
How fast you adopt

Exit
If a vendor weakens a pledge

Coordination held where harm was measurable, a substitute existed and defectors paid a price.

Write down what you will refuse to deploy, and who can say no.

Read the full article at terencekok.com



Terence Kok · AI & Innovation Strategist

SAVE · COMMENT · FOLLOW