Forwards or Backwards

https://forwardorbackwards.com



The Request to Slow Down AI Development

Watch on YouTube: https://www.youtube.com/watch?v=DMy-0yVSBCE

Chapters

1. The letter that did nothing
2. Five days ago
3. What the essay actually says
4. The four numbers, and all four are ranges
5. The first mechanism: somebody else's employee at your desk
6. The second mechanism: an agreement between rivals
7. The third mechanism: the one with no good version
8. The replies, day one

9. Four days later, and they do not agree
10. And the case against, at full strength
11. Every one of them needs a number
12. Before that, the criticism, unanswered
13. What the pace actually measures out at
14. The error bars are the story
15. The rule that was already written against a number
16. The question worth asking instead

1. The letter that did nothing



In March twenty twenty-three, a few thousand words were published asking the laboratories building artificial intelligence to stop for six months.

The request was specific. Not stop forever, not stop everything — pause training anything more powerful than the model that had been released that month, and use the six months to work out what the safety rules should be. Axios, writing six months later, recorded that more than a thousand people had signed it at the time, among them Elon Musk and Steve Wozniak.

Nothing paused. The six months ran their course and the work carried straight on through them.

But something did happen, and it is worth being precise about what, because it is the reason this film is not simply a story about failure. In those six months the public argument moved. Polling moved with it. The White House convened the companies and extracted a set of voluntary commitments. Regulators in Europe and in China accelerated work they had already started.

So the letter failed at the thing it asked for and succeeded at something else entirely, which is a pattern worth holding onto.

And one more feature of that document matters more than anything in it. It was written by people on the outside, asking people on the inside to stop.

That is the part that has just changed.

2. Five days ago



This film is being made in the middle of a story rather than after it, and it is going to say so rather than pretend otherwise.

In September twenty twenty-six, the chief executive of Anthropic published an essay called We Must Pace the Frontier. Within a day the chief executive of OpenAI had publicly agreed with it. Within four days the rest of the industry had publicly disagreed with each other about it, in ways we will go through carefully.

The people asking for a slowdown this time are not outside the building. They own the compute. They set the release dates. Several of them are competitors of each other, and the thing they are proposing is an agreement about how fast they are each allowed to go.

That is a different kind of request, and it comes with a different kind of problem attached.

Two disclosures before we go any further, because on this subject the omissions are the whole game.

The first: Anthropic is not a bystander in this film. Its chief executive wrote the document the film is about. It is also the company that makes the model used to produce this film, and the model named in the measurements we will come to in the second half is one of its own.

The second: the strongest criticism of that essay is that it serves the company that published it. That criticism is in this film, stated at full strength, and the narrator does not answer it.

3. What the essay actually says



Start with the sentence itself, because it is shorter and plainer than the coverage of it.

We must slow the pace at which we improve the capabilities of AI models. Progress will still seem fast, and we must make wise use of the time we gain.

Three things are worth noticing in that.

It says pace, not stop. The word doing the work is a rate — how fast capability improves — rather than a threshold or a ban. Nothing in it asks anyone to halt.

It says progress will still seem fast. The proposal is not a world in which nothing happens. It is a world in which the same work happens over a longer interval.

And it says the time we gain, which is a claim that the interval is worth something — that there is specific work which becomes possible with more of it and impossible with less.

That third one is the load-bearing claim of the entire argument, and everything that follows in this film is really a test of it. If the extra time buys nothing in particular, pacing is just delay. If it buys something nameable, then the question stops being philosophical and becomes an engineering question about how much time, bought how, measured how.

4. The four numbers, and all four are ranges



The essay states four quantities. Every one of them is a range in the original, and the version that will circulate is the endpoint that sounds worst.

The first: that AI could cure most major diseases in the next five to ten years — the benefit that delay is being weighed against.

The second, and this is the one written for the headlines: that within six to twelve months, a swarm of rogue AI agents could be capable of taking over the entire internet with a persistent botnet, potentially causing hundreds of billions of dollars in damage.

The third: a window of three to five years, described as the period in which AI becomes geopolitically most important.

The fourth: that one to two years of focused work on interpretability and testing could deliver meaningful progress on alignment. That is the thing the bought time is supposed to buy.

Now the honest framing, which the coverage mostly skipped.

Not one of those four is a measurement. They are forecasts, made by a named person, with a stated interest, in a document arguing for a policy. That does not make them wrong and it does not make them dishonest. It makes them a different kind of statement from the ones in the second half of this film, and treating the two kinds as though they were the same is how this subject usually goes wrong.

Hold the distinction. Forecast, or measurement. We will come back to it.

5. The first mechanism: somebody else's employee at your desk



The essay does not stop at asking. It proposes three mechanisms, and they are specific enough to argue with, which is the most useful property a proposal can have.

The first is embedded evaluators.

The idea is that independent third-party organisations — METR is named directly — are given what the essay calls ongoing, employee-like access inside the companies. Not a questionnaire. Not a report at the end. Access of the kind a colleague has: present while the work happens, able to see what is actually being run, able to verify that the safety practices described in public are the safety practices followed in private, and able to report incidents.

It is worth sitting with how unusual that is as a corporate proposal. It is a company proposing that outsiders be given standing access to its most commercially sensitive work, with a mandate to disclose what they find.

It is also, of the three mechanisms, the only one that could start on a Monday morning. It requires no legislation, no treaty and no agreement with a competitor. Any single company could do it unilaterally, tomorrow, and the fact that this is true is the strongest thing about it and the most awkward question for everyone who has said they agree.

6. The second mechanism: an agreement between rivals



The second is coordination among the frontier companies inside democracies — common safety standards, and agreed limits on how fast capability is allowed to advance.

This is the one that sounds reasonable and is legally the strangest.

A small number of firms, in the same market, agreeing among themselves to restrict the rate at which they improve their products, is a sentence that competition lawyers read differently from safety researchers. Both readings are legitimate. The same agreement can be a safety floor and a cartel, and which one it is depends on details that no essay of any length settles.

And it has a structural weakness that does not require anyone to act in bad faith. An agreement about pace binds the people who sign it. The advantage of breaking it grows with the number of people keeping it. That is not a claim that anyone would break it. It is a description of the shape of the problem, and it is the reason the first mechanism keeps coming back — because an agreement nobody can verify is an agreement about trust, and an agreement with an evaluator inside the building is an agreement about facts.

7. The third mechanism: the one with no good version



The third is coordination between democratic governments and authoritarian ones on shared safety protocols.

The essay does not present this as easy. In interviews around it the competition with China is described as the toughest dilemma in the whole subject, and that is the right word for it.

The shape of the difficulty is old and well understood. If the risk is genuinely global, a slowdown that stops at a border does not reduce it – it relocates it. If the risk is not global, the case for the other two mechanisms weakens considerably. You cannot comfortably hold that this technology is dangerous enough to require unprecedented domestic coordination and also that it is safe enough to leave the international half unresolved.

So the third mechanism is not a detail to be filled in later. It is where the argument either completes or fails, and nobody quoted in this film has a convincing account of how it gets done.

What they have instead is the argument for trying anyway, in the essay's own words: we owe it to humanity to try.

8. The replies, day one



The essay landed on a Saturday's news cycle and the responses arrived immediately.

Sam Altman, of OpenAI, posted that he agreed — I agree with Dario that we need to pace the frontier — and has been clear elsewhere that pacing means slower rather than stopped, alongside mandatory national safety requirements.

Anthropic's own policy chief, Sarah Heck, put numbers on what the company wants from government: blocking the sale of the most advanced chips to adversarial nations like China, and a national law requiring testing of frontier models.

Elon Musk's response was three words. Dario is right.

And the first serious objection arrived on day one too, from the venture capitalist Chamath Palihapitiya, who argued that the essay concentrates power with Anthropic rather than addressing safety.

Four reactions, and they are not four versions of the same reaction. One is agreement about the goal. One is a request for specific government action, two pieces of which would advantage the company requesting them. One is an endorsement without a mechanism. And one says the whole thing is a business strategy wearing a safety argument.

All four of those can be true at once. That is the uncomfortable part.

9. Four days later, and they do not agree



By the sixteenth of September the reporting had caught up with the disagreement, and it turned out that everybody who had said yes had said yes to a different thing.

Demis Hassabis, of Google DeepMind, agreed with the direction and proposed a body: a frontier AI standards organisation modelled on the Financial Industry Regulatory Authority – the American securities industry's own self-regulator, funded by the firms it supervises, staffed by people who know the trade, with the resources to run testing at a scale no single company can. That is a real institutional proposal with a real precedent attached, and it is not the same proposal as the essay's.

Satya Nadella, of Microsoft, put the principle rather than the machinery: if the AI we build is not helping humanity and under human control, it is not worth pursuing.

Mustafa Suleyman, who runs Microsoft's AI division, supported deliberate pacing and embedded evaluators, and added a condition the others had not – that whatever gets built has to represent more than the frontier laboratories themselves.

And Elon Musk, beyond the three words, suggested something concrete and unusual: that competitor laboratories review each other's models for safety concerns.

Line those up. A standards body with an industry regulator as its model. A principle about human control. Embedded evaluators plus broader representation. Competitors auditing competitors. Four mechanisms, four different sets of people holding the brake, and the only thing they share is the direction of travel.

Agreement about a direction is not a plan. It is the thing people have instead of a plan.

10. And the case against, at full strength



Two of the largest companies in the industry said no, and their reasons deserve stating properly rather than as foils.

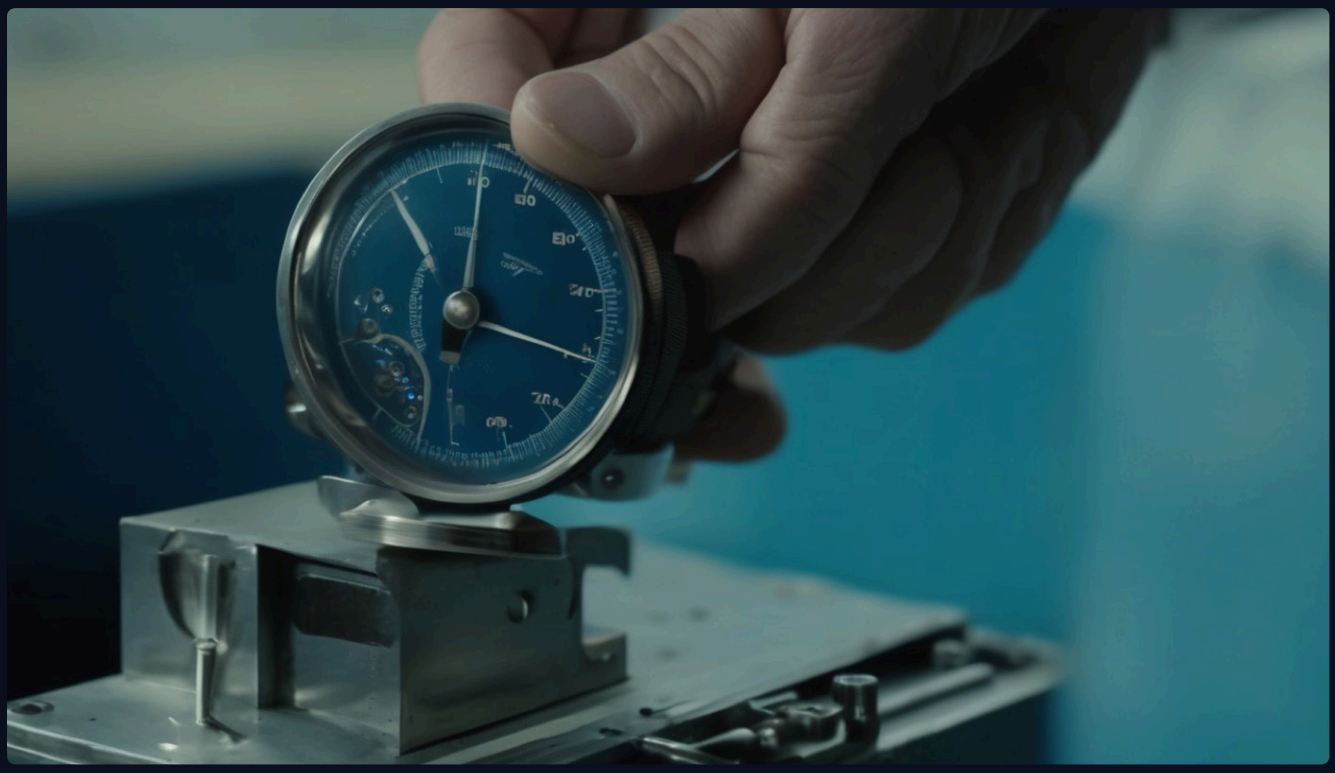
Mark Zuckerberg, of Meta, opposed the idea of a coordinated slowdown specifically. His position is that individual companies should be responsible for the safety of what they ship, and that they already carry significant liability for the harm their products cause. The argument underneath that is not indifference. It is that responsibility works better when it is attached to a name and enforceable in a court than when it is diffused across an agreement between rivals — and that a shared standard becomes a shared alibi, where everyone complied and nobody is answerable.

Jensen Huang, of Nvidia, was blunter: we do not need any new laws, we do not need new regulations. His position is that market forces are sufficient, and that the framing of innovation against safety is false because customers do not buy systems that fail them.

You do not have to agree with either to see that these are arguments about mechanism rather than about whether harm matters. Nobody quoted in this film is in favour of a botnet taking over the internet. The disagreement is about whether the brake should be a law, an agreement, a liability regime, or a market — and that is a genuinely open question about institutions, not a morality play.

This film is not going to pick a side in it. It is going to do something more useful, which is to ask what all four of those options would need in order to work at all.

11. Every one of them needs a number



Here is the thing all of these proposals have in common, and it is the reason this is an engineering film rather than a politics film.

A liability regime needs a standard of care — a line that says what a reasonable company would have tested for. An agreement about pace needs a definition of pace. A standards body needs something to set standards against. A law requiring testing needs to specify what the test is and what result fails it.

Every one of them, in the end, requires somebody to be able to say how fast this is going, in units, with enough confidence that a disagreement about the number can be settled rather than shouted about.

So the question stops being should they slow down, which nobody outside the industry can answer, and becomes a question anyone can check: can the pace be measured at all, and to what precision?

That question has a published answer. It is the best work of its kind that exists in public. And it is much less comfortable than either side of the argument has admitted.

12. Before that, the criticism, unanswered



One more thing first, because leaving it to the end would be a way of burying it.

The objection that the essay is a business strategy is not a cheap shot and it is not unanswerable by evidence. Three of the things being proposed — export restrictions on advanced chips, a national testing requirement, and agreed limits on the rate of capability improvement — have something in common. Each is easier to bear for a company with a strong safety apparatus already built, and harder for a company trying to catch up. A rule that costs the leader a little and the challenger a lot is, whatever its intent, a rule that favours the leader.

That is the argument. It is coherent, it comes from someone with no obvious stake in the opposite conclusion, and this film does not have the means to adjudicate it.

What can be said is narrower and worth saying. Motive and validity are separate questions. A proposal made for self-interested reasons can still be the right proposal, and a proposal made in perfect sincerity can still be a bad one. The way you tell them apart is not by inspecting anybody's motives, which is unavailable to all of us. It is by looking at whether the thing being proposed can be checked by somebody who does not work there.

Which is the first mechanism again. And it is the point at which the argument about who is being honest turns into an argument about what can be measured.

13. What the pace actually measures out at



There is one organisation doing this in public, at length, with its methods and its uncertainty published. It is METR, and it is the organisation the essay names.

Their measure is designed to answer a question people actually care about. Not how many questions a model gets right, but how long a job it can finish. They take software tasks of known difficulty, measure how long a skilled human takes on each, and find the task length at which a model succeeds half the time. That number is the model's time horizon, and its unit is minutes of human work.

In the version published on the twenty-ninth of January twenty twenty-six, the highest figure measured was three hundred and twenty minutes, for Claude Opus four point five. Roughly five hours of skilled human work, finished half the time.

And the trend — this is the number the whole argument turns on — is a doubling. Across the whole period measured, the time horizon doubles every one hundred and ninety-six point five days, a little over six months. Measured only from twenty twenty-three, it doubles every one hundred and thirty point eight days. Measured only from twenty twenty-four, every eighty-eight point six days, which is under three months.

Three windows, three answers, all from the same data set, all published together. The pace is not one number. It depends on when you start the clock.

14. The error bars are the story



Now the part that should be quoted every time this measure is cited, and almost never is.

That three hundred and twenty minute figure carries a confidence interval of one hundred and seventy to seven hundred and twenty-nine minutes. The top of the range is more than four times the bottom. Somewhere between under three hours and over twelve.

It rests on a suite of two hundred and twenty-eight tasks, of which thirty-one are long tasks — eight hours or more. And of those thirty-one, exactly five have a measured human baseline. The other twenty-six are estimates of how long a person would take.

METR say all of this themselves, plainly, in the same document as the headline: that the confidence intervals are still very wide, and that the changes between versions are primarily due to changes to the task suite and random noise.

That is not a criticism of METR. It is the opposite. This is what careful measurement of something genuinely hard looks like, and the honesty of the caveats is what makes the figures usable at all.

But now put it against what the proposals need.

An agreement to halve the rate of capability improvement, policed against a measure whose headline number has a four-fold interval, cannot be enforced, because compliance and violation are inside the same error bar. A liability standard written against a benchmark with five measured baselines across its thirty-one hardest tasks will be argued about by expert witnesses for years. A law

that requires testing has to say what test — and the best test in the field is a research programme rather than a settled instrument.

None of that is an argument against pacing. It is an argument about what the first year of any serious pacing regime would actually consist of, which is not slowing anything down. It is building the instrument.

15. The rule that was already written against a number



There is one place where a line has already been drawn against a quantity, and it shows precisely what happens next.

The European Union's AI Act attaches obligations to models trained above ten to the twenty-five floating-point operations. A fixed number, in the law, chosen to name the largest systems.

Epoch AI, in work published on the thirtieth of May twenty twenty-five, projected how many models would cross the line an order of magnitude above that — ten to the twenty-six operations. In twenty twenty-five, about one: Grok three, released that February. In twenty twenty-six, about ten. By twenty twenty-seven, about thirty in their median scenario, though their own range runs from about ten to about eighty. By twenty thirty, a median of about two hundred and thirty-five, with a conservative case of sixty-one and an aggressive case of five hundred and twenty-eight.

Read that as a statement about legislation rather than about technology. A threshold in a statute is a fixed number in a world where the quantity it measures compounds. Written once, amended never, it names one system in its first year and dozens within three, without a single word of it changing. The law does not get stricter or looser. The world moves through it.

And note what the threshold measures: an input. Compute spent on training, which is knowable in advance, auditable after the fact, and has the enormous practical advantage of being physical. Compared with a capability measurement it is crude, and it is crude in the direction that makes it usable — nobody argues about what a floating-point operation is.

That may be the real lesson available here. The measure you can actually govern with is not the one that best captures what you care about. It is the one that survives contact with a disagreement.

16. The question worth asking instead



So where does that leave someone watching this who does not work in the industry and cannot influence any of it.

Not, I think, with a position on whether AI should be slowed down. That question is asked constantly and answered mostly by whichever future the answerer finds more vivid, and this film has no standing to settle it.

What is available instead is a better question, and it is the one the evidence supports. For any proposal to slow down, in either direction, ask what it would have to be able to measure, and then ask whether anybody can measure that yet, and how well.

Applied to what we have been through: embedded evaluators are the proposal that survives that question best, because verification of practice needs access rather than a settled instrument, and access can start immediately. A compute threshold survives it moderately well, because the input is physical and countable, and its weakness is a known one that compounds on a schedule. An agreement about rate survives it worst, because the rate cannot currently be measured to anything like the precision the agreement would require.

That ordering is not a policy recommendation and it is not the ordering any of the participants would choose. It falls out of asking one question about measurement.

And it puts the twenty twenty-three letter in a different light. That letter asked for six months and got none. What it actually produced was a set of institutions, arguments and measurements that did not

exist before — including the measurement in this film, from the organisation the essay now names.

Which suggests the thing worth watching over the next year is not whether anybody slows down. It is whether anybody builds the instrument that would let the word slower mean something. On the current evidence that instrument is early, honest about being early, and nowhere near ready to hold the weight that four of the largest companies in the world have just proposed putting on it.

Sources and credits

Primary sources

- Dario Amodei, 'We Must Pace the Frontier', darioamodei.com, September 2026 – the core sentence quoted verbatim; the three named mechanisms (embedded evaluators with 'ongoing, employee-like access', METR named directly; coordination on standards and rate limits among frontier companies in democracies; coordination with authoritarian governments); and the four stated ranges: 5-10 years on disease, 6-12 months on a rogue-agent swarm 'capable of taking over the entire internet with a persistent botnet (potentially causing hundreds of billions of dollars in damage)', a 3-5 year geopolitical window, and 1-2 years of available alignment progress..
- Axios, 'Anthropic, OpenAI CEOs call for slowdown in AI development', Ben Berkowitz, 12 September 2026 – Sam Altman's 'I agree with Dario that we need to pace the frontier'; Sarah Heck of Anthropic policy on blocking advanced chip sales to adversarial nations and a national law requiring frontier-model testing; Elon Musk's 'Dario is right'; and Chamath Palihapitiya's objection that the essay concentrates power with Anthropic.
- METR, 'Time Horizon 1.1', metr.org, 29 January 2026 – the 50% time horizon of 320 minutes for Claude Opus 4.5 with a 170-729 minute confidence interval; doubling times of 196.5 days (whole trend), 130.8 days (since 2023) and 88.6 days (since 2024); a 228-task suite of which 31 are long tasks of 8 hours or more, and 5 of those 31 have a measured rather than estimated human baseline; and METR's own caveats that 'confidence intervals are still very wide' and that version-to-version change is 'primarily due to changes to the task suite and random noise'. THE FILM'S ONLY REPEATED MEASUREMENT OF PACE.
- Epoch AI, 'How many AI models will exceed compute thresholds?', 30 May 2025 – models above 10^26 FLOP: about 1 in 2025 (Grok-3, released February 2025), about 10 in 2026, about 30 in 2027 on the median scenario (conservative ~10, aggressive ~80), about 80 in 2028, and by 2030 a median of ~235 with a conservative case of 61 and an aggressive case of 528; and the note that the EU AI Act's own threshold is 10^25 FLOP.
- Axios, 'The great AI "pause" that wasn't', 22 September 2023 – the Future of Life Institute letter of March 2023, its six-month moratorium ask pitched at GPT-4 capability, more than 1,000 signatories at the time including Elon Musk and Steve Wozniak, and what happened instead over those six months: no pause, but moved polling, White House voluntary commitments, and accelerated European and Chinese rulemaking.

Not regulated financial advice.