Could AI Destroy Humanity?

Can We Afford to Slow Down?

By Dr Peter Fields | First posted: 5 November 2026

THE 2-MINUTE SUMMARY

Should powerful AI development be slowed down? A pause may allow researchers and regulators to understand dangerous capabilities and strengthen defences. Yet delay can also postpone beneficial tools, including those used to protect hospitals, businesses and essential infrastructure. Attackers and defenders do not necessarily stop together.

This article distinguishes research, training, release and deployment, rather than treating all AI development as one activity. It examines arguments for targeted pauses, the costs of waiting, cyber-security evidence and international competition. Its conclusion is conditional: any delay should identify the specific risk, explain what will improve during the interval and show why the resulting protection is likely to outweigh the costs.

Suppose an AI system finds a security flaw in a hospital's software. It could help protect patients or help an extortionist shut the hospital down. Asking whether to develop that capability faster or more slowly does not, on its own, tell us which outcome becomes more likely.

That is the difficulty behind the title. Slowing AI might give us time to understand dangerous behaviour and strengthen our defences. It might also delay the tools needed to protect us. Meanwhile, criminals and hostile states are unlikely to suspend their activities while we settle the argument.

My starting point is that delay needs a purpose. We should ask what will change during it, who will keep moving, and whether we will be better protected when it ends. The same questions belong on the desk of anyone demanding faster development.

What exactly are we slowing down?

“AI development” bundles together several decisions. Researchers can investigate how a system works. A company can train a more capable model, release it to users, or make copies available for others to run independently. A bank can give an existing model permission to suggest transactions, approve them, or execute them. These are different choices.

Restricting the next generation of frontier models, the most capable models being developed, need not stop useful work with existing ones. Delaying a public release need not prevent controlled testing. We could spend longer checking a powerful system while moving faster to repair weaknesses in essential services.

The distinctions have limits. Research can itself create dangerous capabilities. An experimental agent with access to live networks can cause harm before it becomes a product. Existing models may also be unsafe for a proposed task. But a policy aimed at one problem should explain why it restricts the others.

Amodei’s strongest argument

Dario Amodei’s September 2026 essay, We Must Pace the Frontier, deserves a more careful reading than the word “pause” invites. He explicitly distinguishes pacing from halting model training or technical progress. His concern is that capability improvement is outrunning the work needed to keep systems under control.[1]

A particular concern is the feedback loop: AI helping researchers build the next generation of AI, which then helps build its successor. Amodei argues that this is accelerating progress.[1] That deserves attention without being treated as proof of an unstoppable intelligence explosion. It also explains why delaying a public launch might be insufficient if internal development continues at full speed.

He identifies practical uses for additional time: more reliable operations, better training to produce the intended behaviour, improved methods for understanding what happens inside models, and stronger testing. He also proposes embedded external evaluators with continuing access to the development process and rights to publish adverse findings, subject to specified redactions.[1]

The argument has force. If a system's failures are poorly understood, making its successor more capable may increase the consequences before anyone has repaired the cause. Commercial pressure can make it difficult for one developer to spend longer on checks while competitors release their products.

Amodei believes another year or two could substantially improve safety. That is a judgement about what research and engineering might achieve, not an established result. Yet dismissing it because the outcome is uncertain would be too easy. Almost any serious programme to reduce an emerging risk must begin before its success is guaranteed.

A pause with a stated purpose

There is already a concrete example. On 18 August 2026, OpenAI reported a two-week pause in reinforcement learning, a stage of training that rewards selected behaviour, on its latest models intended for deployment. It said its largest planned frontier run of that kind remained on hold while smaller-scale work continued. The stated purpose was to strengthen security, monitoring and evidence that models would behave as intended.[2]

This was a selective pause in training, on the company's account. The August statement does not establish that the hold continued into November, or independently demonstrate that the safeguards worked.

In September, Sam Altman also endorsed Amodei's pacing proposal and promised embedded external evaluators.[3] A commitment deserves scrutiny. It does not establish effective independent access or a binding limit on development.

These examples bring the question into focus. Which failures will a delay address? What access will independent researchers receive, what resources will be committed, and what evidence will count as progress? A programme that resolves a dangerous failure could justify taking longer. A year that produces little beyond another statement of principles would be harder to defend.

There must also be room to discover that the original deadline was wrong. A review date should require a decision on the evidence, rather than guarantee release because the calendar has run out.

The cost of waiting is real

If a tool can reliably prevent harm, withholding it can have consequences. Safety includes people who remain exposed while an effective protection waits for approval.

Cyber security offers a concrete example. In DARPA's 2025 AI Cyber Challenge, autonomous systems combined AI with security techniques to find and patch flaws in real software. DARPA reports that the finalists collectively found 54 of 63 deliberately inserted vulnerabilities and patched 43. They also found 18 genuine vulnerabilities that had not been inserted for the competition and provided 11 patches for those.[4]

Those results show useful capability under competition conditions. They do not establish that a generated patch is ready to install without checking. Nevertheless, they give the case for defensive AI more substance than a promise about what artificial general intelligence might eventually do.

The distinction matters when discussing other hoped-for benefits too. A plausible contribution to medical research is a reason to investigate. It is not yet a demonstrated cure, and a forecast of future cures cannot simply be converted into a count of people killed by a proposed restriction.

The costs of moving too quickly and of waiting both need scrutiny. Neither should acquire a fictitious precision just because a number makes the argument more dramatic.

Attackers and defenders do not move together

The National Cyber Security Centre's annual review, published on 3 November 2026, assesses that AI will increase the speed, scale and efficiency of hostile cyber activity. It warns that AI-assisted vulnerability discovery could give attackers more opportunities to strike before repairs are installed. It also anticipates a widening gap between organisations able to adopt AI for defence and those unable to adapt.[5]

The same review also judges that AI could ultimately be a net positive for cyber security, provided urgent action is taken to realise its defensive potential.[5] Its warning is a case for doing that work, not a conclusion that defence is futile.

In a September article, the NCSC's Dave Chismon explains the asymmetry. An attacker may succeed by making software fail. A defender must protect it while keeping the business running, securing budgets and obtaining permission for changes. A repair that shuts down an essential service can itself cause harm. His argument supports carefully bounded automation; it does not reject AI-assisted defence.[6]

A national laboratory might demonstrate an impressive protective tool while a small hospital lacks the staff, money or suitable software to use it. The existence of the tool does not mean the hospital is protected.

Defensive policy therefore needs to address adoption as well as invention. Testing repairs, replacing insecure older systems and helping under-resourced operators may matter

more immediately than another improvement in a model's benchmark score. Maintaining basic security should not wait for the next model.

Geek Box

The time between finding a flaw and fixing it

Consider an invented example. A flaw becomes known on Monday. Defenders can install a checked repair on Friday; an attacker can build a working exploit by Wednesday. There is a period when the system remains vulnerable and the attacker can act.

If AI brings the repair forward to Tuesday, the position improves. If it brings the exploit forward to Monday while the repair still waits until Friday, it worsens. Actual risks also depend on access, detection and whether the attack succeeds. The example illustrates why faster AI is not, by itself, a measure of greater safety.

This is also why a restriction on new frontier training and an investment in defensive deployment could be consistent parts of the same policy. Conversely, obstructing useful defensive tools while leaving comparable attacking capabilities readily available would need a compelling justification.

Criminals will not obey a pause

A prohibition does not make existing knowledge disappear. Nor can a developer assume that independently held copies of released model weights, the numerical parameters used to run a model, will vanish when it withdraws its own service. In its March 2026 briefing, the NCSC explains that publicly available weights can be modified and safeguards removed.[7]

That makes the moment of release significant. Once a capability is widely available, our options change. Preventing its initial distribution may have been possible; retrieving every copy is a different proposition. This strengthens the case for examining releases beforehand, while limiting what a later ban could achieve.

There is a cost to restricting open weights too. They let independent researchers and less well-funded developers inspect, adapt and build on models, as the International AI Safety Report recognises.[8] Closing that route can increase dependence on a few suppliers. The case for restricting a release should therefore identify its dangerous capabilities and weigh the research and competition costs, rather than treating openness itself as the offence.

Securing models against theft, restricting access to sensitive systems and improving the resilience of legitimate operators can make an attack harder even when the attacker ignores every rule. Incomplete enforcement does not make those precautions redundant.

North Korea should not become a shortcut in the reasoning. The question is whether a hostile actor can obtain and use a harmful capability. That is different from whether it can

build the world's most advanced model independently. Maintaining our ability to defend against that actor matters; it does not follow that every new capability must immediately become available to everyone.

What if China keeps going?

International competition presents a harder version of the problem. A restraint that leaves one group of countries substantially behind another could weaken its security and its ability to influence how the technology is used. That possibility cannot be dismissed as a commercial excuse.

As Part 3 explained, China is not an unregulated exception. Its 2023 interim measures govern generative AI services supplied to the mainland public and include content controls and specified security-assessment requirements. Their scope does not establish equivalent restraints on all frontier research.[9] The question is what each country actually constrains, rather than which has more rules.

Amodei acknowledges the strategic objection. He argues that pacing among democracies is limited by their lead over China, and proposes tighter controls on advanced chips and stronger protection against model theft. He treats comprehensive global pacing as much harder to achieve than agreements on particular dangerous uses or testing.[1]

Those proposals do not prove that a lead can be preserved. Restrictions may fail to bind a rival; agreements may be difficult to verify. Countries may lead in different capabilities, making the language of a single race misleading.

Before supporting a broad international slowdown, I would require a credible account of compliance and strategic consequences. Narrower agreements may still help. Rivals can share an interest in preventing attacks that damage both without trusting each other with unrestricted access to their research.

Uncontrolled competition has costs too. Being first to deploy a system you cannot reliably control is a doubtful security advantage. A lead matters through the useful, dependable capability it provides.

Defence needs limits of its own

Calling a system defensive does not make its actions harmless. The NCSC's March briefing warns that automated responses can cause service disruption, data loss or operational failure, potentially exceeding the impact of the original attack.[7] A tool authorised to isolate suspicious computers might also isolate computers essential to keeping a service running.

This is where the argument returns to Part 2. We need to examine what a system can do, what access it has and how failures can be contained. A recommendation that a competent person checks differs from a system changing live infrastructure faster than anyone can intervene.

I would also expect safety checks to continue after deployment, and to be revisited when training, tools or access change. The system should not control its own monitoring or the decision to stop it. Operators need the authority to intervene and a workable recovery plan. A successful test before launch is useful evidence, not a permanent certificate of safety.

Nor does adding a human approval button settle the matter. If the operator lacks the information, time or expertise to challenge a recommendation, oversight may become ceremonial. Preserving that competence is part of maintaining control. The wider question of how people acquire experience when AI takes over their early work deserves its own discussion.

What would justify the delay?

Cyber security makes the dilemma tangible, but cannot settle every AI risk. Biological misuse and loss of control have different pathways and safeguards, as the International AI Safety Report explains.[8] Evidence that a model repairs software cannot, by itself, justify releasing every other capability it possesses.

My test for a proposed slowdown would begin with the dangerous activity being restrained and the protection the delay is intended to achieve. There should be a funded programme of work, people able to examine its results independently, and a timetable for reviewing whether the restriction is helping. It should also identify which useful activities can continue.

Uncertainty can itself warrant restraint where there is a credible route to severe harm, weak containment and little ability to reverse the consequences. We should examine the evidence for the steps connecting capability to harm. We need not wait for the whole sequence to end in catastrophe before acting. But an indefinite restriction with no account of how conditions might improve deserves challenge as well.

The developer pressing ahead owes an explanation too. What supports its claim that the system is ready? Who will bear the loss if it is wrong? “Our competitors are doing it” explains an incentive. It cannot establish that the proposed action is safe.

Can we afford to slow down? Sometimes, particularly where a delay can produce safeguards before an irreversible release. Elsewhere, delaying tested protection could leave people exposed. The answer depends on what we restrain and what we do with the time.

I would support taking longer to understand and contain a dangerous capability while investing urgently in protection against it. We must show that those choices improve our position. A slower calendar offers no protection on its own; neither does a faster model.

Sources and further reading

1. Dario Amodei, We Must Pace the Frontier, September 2026. See Why Pace?, Embedded Evaluators, Pacing Within Democracies and Global Pacing. His estimates of what additional time could achieve, and of the effectiveness of strategic controls, are proposals and judgements, not established outcomes. Read source

2. OpenAI, Pacing Model Development in an Era of Cyber-Critical Capabilities, 18 August 2026. See the opening account of the two-week reinforcement-learning pause and the largest planned frontier RL run, and the sections on research security, monitoring and alignment. These are the company's statements as at that date; they do not independently establish effectiveness or the position in November. Read source

3. Anthony Ha, Anthropic CEO Outlines Plan to Slow AI Development, TechCrunch, 12 September 2026, updated with comments from Sam Altman and Elon Musk. Reports Altman's endorsement and commitment to embedded evaluators. A public commitment is distinct from verified implementation. Read source

4. DARPA, AI Cyber Challenge Marks Pivotal Inflection Point for Cyber Defense, 8 August 2025, subsequently corrected. See Deep Dive: 54 of 63 deliberately inserted vulnerabilities found, 43 patched; 18 non-synthetic vulnerabilities found and 11 patches provided for those. This article uses the reported counts. Competition performance is not a guarantee of safe deployment. Read source

5. National Cyber Security Centre, Annual Review 2026, published 3 November 2026. See The Impact of AI on Cyber Threat, printed pp. 21–23, on accelerating hostile activity and unequal defensive adoption; and printed p. 61 on the potential net benefit and the urgent action required. These are assessments and forecasts, not measured probabilities of catastrophe. Read source · Full annual review and PDF

6. Dave Chismon, One Does Not Simply Defend Agentically, NCSC, 21 September 2026. Explains the organisational constraints on defenders, the risks of automated changes and a framework for bounded defensive action. It does not oppose defensive AI in general. Read source

7. Paul J and Alan Steer, Why Cyber Defenders Need to Be Ready for Frontier AI, NCSC, 30 March 2026. See How Frontier AI Capabilities in Cyber Operations Are Evolving, on modifying publicly available model weights and removing safeguards, and Automating Mitigation and Response, on risks from automated defensive actions. Its dated model-capability evaluations are not treated here as a description of all capabilities in November 2026. Read source

8. International AI Safety Report 2026, Extended Summary for Policymakers, 3 February 2026. See sections 2.1.4 (biological and chemical risks), 2.2.2 (loss of control) and 3.4 (open-weight models). The report distinguishes benefits, possible harms and evidential uncertainty; it does not endorse a particular regulatory policy. Read source

9. Cyberspace Administration of China and six other authorities, Interim Measures for the Management of Generative Artificial Intelligence Services, Order No. 15, published 13 July 2023, effective 15 August 2023. See articles 2, 4, 17 and 24. Official Chinese text, paraphrased in English. The measures' scope is not evidence that equivalent constraints apply to all frontier research. Read source