Could AI Destroy Humanity?

Who benefits from regulating AI?

By Dr Peter Fields | First posted: 4 November 2026

THE 2-MINUTE SUMMARY

Calls for AI regulation should be judged by the rules proposed and their likely consequences, not simply by the identity of the companies supporting them. Safety obligations can protect the public. They can also impose compliance costs, restrict competition or favour established developers. These effects are not mutually exclusive.

This article considers who bears AI risks, why major companies might welcome regulation, how restrictions could affect smaller competitors and innovation, and what happens when other countries or criminals do not follow the same rules. The central distinction is between proportionate safeguards directed at identifiable harms and measures whose benefits, costs and enforcement have not been adequately established.

When an AI company asks to be regulated, my first question is: what rules is it asking for? A requirement to report a serious failure, a licence to train a model and a prohibition on a particular use are very different interventions. Each can produce different winners.

In June 2023, responding to questions following his US Senate appearance, Sam Altman supported registration, disclosure and licensing requirements for future, highly capable general-purpose models. He also acknowledged the difficulty of defining their scope and avoiding unnecessary burdens on smaller firms.[1] In June 2026, Anthropic proposed independent evaluation and government powers to block or deter dangerous deployments in its Advanced AI Framework, with its principal requirements targeted at large developers of powerful models.[2]

There are reasonable grounds for asking whether such proposals protect the public, protect the companies proposing them, or do some of both. We should examine their effects before deciding what to make of their supporters.

Who bears the risk

Motor vehicles and powered flight expose people to harm. Their usefulness has given us reasons to manage those dangers rather than abandon the technologies. That observation belongs in the AI debate, although it cannot tell us how much risk to accept.

AI could help discover medicines, improve access to expertise and secure vulnerable software. Delaying useful applications has a cost. Equally, allowing a technology to advance does not require accepting every proposed use of it, at every speed and with every degree of autonomy.

A familiar distinction matters here. Someone might willingly try an experimental tool on their own documents. Connecting an autonomous agent to a hospital's operational systems exposes people who may have had no part in choosing it. The expected benefit to the purchaser cannot, on its own, justify the risk imposed on patients.

In Part 2, I argued that consequential authority deserves stronger controls. Regulation can make those controls less dependent on whether a particular organisation finds them commercially convenient. It can also require information that customers and affected people cannot obtain for themselves.

The assessment must include the likely harm, its scale and reversibility, the benefits forgone through delay, and the practical alternatives. Uncertainty makes that judgement

harder. It does not justify either treating every alarming possibility as decisive or assigning it a value of zero.

Why a company might welcome rules

There is a straightforward commercial difficulty behind calls for common standards. A company that spends longer testing a product may lose business to a competitor that releases sooner. If testing genuinely reduces serious risk, shared requirements could remove part of the reward for cutting corners.

Rules can also give businesses a more predictable basis for investment. Customers may be more willing to adopt AI if they can establish what testing has occurred and who is accountable. A developer could therefore gain from regulation because its customers gain confidence in the product.

Other advantages deserve attention. Helping to define the standard may favour a company's existing methods. A demanding approval process can make entry harder for competitors. Presenting oneself as a responsible participant can enhance a company's reputation. None of these possibilities establishes why a particular executive supports a particular proposal.

Sincere concern and commercial self-interest can coexist. A useful safeguard does not become useless because a company benefits from it. Nor does a company's expertise entitle it to decide what everyone else must do.

The cost of getting through the door

The competition concern has a factual foundation beyond speculation about personalities. In 2024, the Competition and Markets Authority warned that powerful firms could use control over essential inputs, routes to customers and partnerships to reinforce their positions in AI markets. It also recognised that those partnerships could support innovation.[3] That warning does not establish that any particular AI safety proposal was an attempt at regulatory capture.

Regulatory capture describes a situation in which public oversight comes to serve the interests of those being regulated. It need not begin with a secret agreement. Dependence on industry expertise, repeated access to policymakers and standards built around established practices can all warrant scrutiny.

An expensive compliance process can reinforce an existing advantage even when everyone faces the same rule. A new entrant may have a better or safer product but lack the resources to demonstrate compliance in the prescribed way.

Geek Box

The same bill can impose a different burden

Suppose an approval process costs £1 million. Against annual revenue of £100 million, that is 1%. Against £5 million, it is 20%. These figures are invented to illustrate a fixed cost, not estimates of actual AI compliance spending.

The calculation does not show whether the requirement is justified. Some expensive tests may be essential. It shows why "the same rules for everyone" does not settle whether firms can compete fairly, or whether the same protection could be achieved less expensively.

I would therefore want an explanation of why a proposed test is necessary, whether smaller firms can access it, and whether there are equivalent ways to demonstrate safety. Size-based exemptions may be justified, but their scope must remain responsive to what a system can actually do.

Anthropic's own proposal illustrates the trade-off. Its main duties apply only when a model exceeds a threshold for the amount of computation used in training and its developer exceeds a financial threshold. That limits compliance costs for smaller firms, but could leave a future dangerous model outside these particular duties if it came from a developer below the financial thresholds. The framework recognises that cheaper training may require thresholds based on what a model can do, and proposes annual review. I would want a clear route for bringing such a model within scope, alongside proportionate ways for smaller firms to demonstrate safety.[2]

Competition itself can matter to resilience. If many organisations depend on one supplier, a failure there can affect them together. Diversity can reduce some shared dependencies, although having more suppliers does not automatically make each product safer.

Regulation can slow development

Approval queues, uncertain obligations and the prospect of substantial penalties can deter investment or delay a release. It would be unconvincing to deny that. Sometimes delay is the intended protection, for example while a dangerous failure is investigated. In other cases it may buy little beyond completed forms.

The European Union's 2026 changes to its AI Act illustrate the implementation problem. Among other changes, they postponed application of specified high-risk rules to December 2027 and August 2028, while also simplifying some requirements. Other parts of the framework already apply.[4] This is evidence that the practical timetable matters, not a verdict that the Act as a whole helps or hinders innovation.

We should also distinguish research, model training and deployment. Investigating a dangerous capability under controlled conditions is different from making it widely available. Restricting a particular deployment need not prohibit work on a safer version. Conversely, research involving networked agents may itself need containment; the word "research" cannot settle the question.

For any proposed restriction, I would ask what evidence could permit it to be lifted. A rule with no intelligible route to approval may discourage an improvement as readily as a dangerous product. A rule that accepts reassuring paperwork without examining behaviour may achieve very little.

What if other countries keep going

The international objection is substantial. If one group of countries imposes effective restraints while another develops comparable capabilities without them, the balance of power could change. An agreement is worth less if nobody can establish whether the other participants are observing it.

But “the West regulates and China does not” is an inaccurate starting point. China’s interim measures for generative AI services took effect in August 2023. They include content controls and, for services with public-opinion or social-mobilisation capabilities, security assessments and algorithm filings. Their scope centres on services offered to the public in mainland China and excludes specified development and applications not offered to that public.[5]

Those measures do not establish that China imposes equivalent constraints on frontier research or addresses the same safety concerns. Comparing regulatory regimes requires examining what they cover and how they are enforced. Counting rules tells us little about their effect on strategic capability.

Dario Amodei recognises the problem. In his September 2026 proposal to pace frontier development, he distinguishes coordination among democracies from the harder task of global coordination, and explicitly warns about losing a lead to China. His proposed response includes verification, stronger security and restrictions on access to advanced chips.[6] Whether those measures would secure enough cooperation or additional time remains a separate question.

His proposal also asks the US government to enable certain safety discussions through a narrow waiver from US competition (antitrust) law. The wider plan envisages common standards and limits on unchecked progress among frontier developers in democratic countries.[6] That might reduce pressure to cut corners. It could also allow established firms to influence the terms on which competitors advance. I would want the scope, oversight and treatment of new entrants made explicit before accepting that the public benefit outweighs the competition risk.

His proposal should therefore be tested against the geopolitical objection, rather than dismissed as if he had overlooked it. I would need stronger evidence for a broad, coordinated slowing of capability development than for a requirement to preserve incident records or secure a deployed system. Those policies impose different costs and demand different kinds of cooperation.

Criminals will not wait for permission

North Korea is a useful test of the argument, provided we keep the claim precise. Its cyber operations are well documented, including theft and attempts to evade sanctions.[7] That does not establish that it could independently develop frontier AI. A hostile actor can nevertheless cause serious damage using capabilities developed elsewhere.

A legal prohibition would not make existing models, technical knowledge or criminal demand disappear. We should expect attempts to obtain models illicitly, remove safeguards or offer services beyond effective enforcement. Criminals do not need an AI that shares their motives, or exclusive possession of the most advanced system.

That limits what regulation can achieve. It does not make every rule pointless. Security requirements can make theft harder. Restrictions on consequential access can limit damage. Enforcing duties on legitimate operators can reduce the number of poorly protected systems available for attack, even where the attacker ignores the law.

Maintaining defensive capability is therefore part of the safety argument. In its annual review published on 3 November 2026, the National Cyber Security Centre assesses that AI will increase the speed and scale of hostile activity. It also warns of a widening gap between organisations able to use AI for defence and those unable to adapt.[8]

Anthropic's framework also proposes investment in societal resilience: stronger underlying software, support for under-resourced infrastructure operators, and safeguards for wider access to defensive tools.[2] Those measures deserve separate assessment from restrictions on developers. A policy can strengthen our defences as well as constrain a product.

A policy that obstructs defenders while doing little to constrain attackers deserves criticism. Yet that does not establish that releasing every new capability immediately improves our security. Access arrangements, safeguards and defensive deployment matter. Faster development may help both sides; the practical question is which measures improve the position of those protecting essential systems.

What would useful regulation require

The strongest starting point, in my view, is to require evidence about consequential systems and make it possible to act on that evidence. The model developer and the organisation deploying it control different parts of the risk. Both need responsibilities suited to the choices they can actually make.

Independent evaluators need meaningful access and the freedom to report adverse findings. Their competence, funding and conflicts also deserve scrutiny. Replacing a developer's assertion with a consultant's certificate does not, by itself, establish safety. Technical claims must remain open to challenge, including by people outside the largest companies.

There is a concrete UK example of the narrower approach. Government proposals updated on 3 November envisage agentic-AI resilience requirements, notification of affected organisations, preservation of evidence and post-incident review within the cyber resilience regime. These remain measures intended for future implementation after consultation and parliamentary approval, rather than a comprehensive AI law already in force.[9]

Such requirements would not answer every frontier risk. They would help establish what happened when control failed and allow others to learn from it. That is a more testable objective than a general promise to ensure responsible AI.

We should ask the same practical questions of regulators that we ask of developers. Do they have the expertise and resources to investigate? Can decisions be challenged? Are the rules reducing a demonstrated danger, and are they being revised when the evidence changes? Passing legislation is the beginning of that work.

Who should benefit

The answer to the title is not one group. Customers may gain protection. Established companies may gain trust, predictable standards and a competitive advantage. Smaller firms may benefit from clearer expectations while struggling with the cost. Governments acquire powers whose use also requires limits.

My test would be whether a rule gives protection proportionate to its costs, without unnecessarily concentrating control over the technology. That requires examining safety and competition together. We should be wary both of firms seeking immunity in the name of innovation and of rules that make their position harder to challenge without a corresponding public benefit.

There is also a danger beyond the immediate failure of an autonomous system. Organisations can become dependent on suppliers, and people can become accustomed to accepting answers they no longer know how to assess. Regulation cannot preserve judgement on our behalf. Nor would a safety certificate tell an employer whether replacing junior experience with automation is a sound long-term decision. That deserves its own discussion.

For this series, the distinction is simpler. We can require stronger evidence before granting an AI consequential authority without assuming that intelligence produces hatred, or that every delay is a victory for safety. The obligation should become more demanding as the potential consequences grow. The route to meeting it should remain open to a challenger with a better answer.

Sources and further reading

1. Sam Altman, Questions for the Record, US Senate Committee on the Judiciary, 22 June 2023, responding to Senator Kennedy, question 1. Supports registration, disclosure and licensing for future highly capable foundation models, and identifies the burden on smaller firms as a design issue. Read source

2. Anthropic, Advanced AI Framework, June 2026, a proposal written primarily for the US federal government, not enacted law. See pp. 3 – 5 on scope, review and other applicable law, and pp. 13 – 19 on societal resilience, especially cyber resilience at pp. 16 – 18. Its proposed developer duties require both more than 10^{25} training FLOPs (floating-point operations) and either more than $500 million in annual AI-derived revenue or more than $1 billion per year in AI research and development spending. It anticipates possible capability-based thresholds and at least annual review of the criteria. It also seeks to preserve other state law and states that compliance should not itself confer immunity under otherwise applicable state law. Read source

3. Competition and Markets Authority, CMA Outlines Growing Concerns in Markets for AI Foundation Models, press release, 11 April 2024, accompanying its update paper. Identifies competition risks from control of inputs, routes to market and partnerships; also recognises potentially beneficial partnerships. Cited for those structural concerns, rather than as a finding that a specific safety regulation constitutes regulatory capture. Read source

4. European Commission, AI Act, Application Timeline and How Has the Commission Simplified the Implementation of the AI Act?, accessed 4 November 2026. Records the 2026 amendments, their entry into force on 27 July, and revised application dates of 2 December 2027 and 2 August 2028 for specified high-risk rules. Other obligations have earlier application dates. Read source

5. 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; the article paraphrases its scope and requirements in English. Read source

6. Dario Amodei, We Must Pace the Frontier, September 2026. See Pacing Within Democracies, Global Pacing and footnote 1, including the proposed narrow antitrust waiver for certain safety conversations and the proposal to strengthen security against model-weight theft. Cited for his proposed approach and acknowledgement of geopolitical and verification difficulties, not as proof that the proposed measures would work or that a waiver has been granted. Read source

7. National Cyber Security Centre, Annual Review 2024, Chapter 01, Democratic People's Republic of Korea. Documents North Korean revenue-raising, sanctions-circumvention and intelligence-related cyber activity. This is evidence about cyber operations, not a finding of independent frontier-AI development. Read source

8. National Cyber Security Centre, Annual Review 2026, The Impact of AI on Cyber Threat, 3 November 2026. See assessments of attack speed, scale and efficiency, and the widening divide in defensive adoption. These are NCSC assessments, not measured probabilities of catastrophic outcomes. Read source

9. UK Government, Agentic AI Resilience Measures, updated 3 November 2026, especially How Are We Going to Do It? and Implementation. Describes intended measures under the Cyber Security and Resilience (Network and Information Systems) Bill regime, subject to consultation and the affirmative parliamentary procedure. Read source