Could AI Destroy Humanity?

Part 1: A 10% Chance of the End of Humanity?

Peter Fields · First posted: 3 November 2026

THE 2-MINUTE SUMMARY

Before asking whether AI has a 10% chance of destroying humanity, we need to explain what could make such an outcome possible. A system pursuing a human-assigned task might find that extra resources, continued operation or reduced oversight would help it succeed. But an instrumental advantage is not the same as a desire to rule, and mathematics does not show that every capable AI will seek power.

This article examines Hinton's warning, the fragmented reality of AI systems, and the separate steps between unauthorised action, loss of human control and extinction. Only then does it consider the estimates in the headlines. The question is not merely how alarming a percentage sounds, but which mechanisms, evidence and assumptions would justify it.

My instinctive response to headlines about AI destroying humanity has usually been scepticism. The leap from a more capable computer to the end of our species often seems to happen somewhere between the interview and the headline, with rather little explanation in between.

Yet some of the people raising the alarm helped create the technology. Geoffrey Hinton is not simply a commentator who has discovered that frightening predictions attract attention. His work helped lay the foundations of modern AI. Dario Amodei runs Anthropic, the company behind Claude. Their concerns deserve examination, even where their conclusions remain open to challenge.

I want to understand what they think could happen, why they think it could happen, and how far the evidence takes us. Before examining the startling percentages, we should ask a more basic question: how could the catastrophe actually happen?

How could AI actually destroy humanity?

Before attaching a percentage to human extinction, we need to explain how it could happen. Saying that future AI could become more intelligent than humans is not enough. Intelligence is not a motive; nor does it confer access to electricity grids, financial systems or the physical

means of causing harm. What objective would produce dangerous behaviour? What opportunities would make it possible? Why could people not intervene?

Hinton offered a memorable analogy in a January 2025 interview. A parent in a hurry may take over tying a young child's shoelaces, not out of hostility, but because doing the job directly is easier. Hinton wondered whether a far more capable AI might likewise take control of a task.[9] It is an illustration of his concern, not evidence that a more intelligent system will inevitably behave like the parent.

We should distinguish two possibilities. People might deliberately use AI to inflict harm. In that case the hostile objective comes from humans. The more difficult question is whether a system pursuing a legitimate task could generate harmful strategies that nobody instructed it to pursue. Neither possibility requires consciousness, boredom, resentment or a machine suddenly deciding to attack humanity.

Why would pursuing a task lead to seeking control?

Imagine an autonomous agent responsible for maintaining an industrial process. More information, computing capacity or time might help it succeed. Interruptions might obstruct it. In some circumstances, it could attempt to obtain additional resources, postpone shutdown or conceal a difficulty. Such steps are sometimes called instrumental goals: means that appear useful for achieving a separate objective.

But an incentive to finish the job is not automatically an incentive to control humans. The agent might request resources, find a more efficient solution or report that the task cannot be completed safely. To depart from those alternatives, it would need some propensity to disregard constraints, together with the ability and opportunity to act. An instrumental advantage, an attempted action and a successful takeover are three different things.

GEEK BOX

Does mathematics prove that advanced AI will seek power?

In Optimal Policies Tend to Seek Power (2021), Alexander Matt Turner and colleagues examined idealised decision-making systems. Under specified mathematical conditions, preserving a wider range of future options is optimal for many reward functions.[10] This gives a rigorous explanation for some power-seeking incentives without assuming ambition or a human personality.

The result is conditional. It is not proof that every language model will seek power, that every advanced system will resist shutdown, or that controlling humanity is the best way to perform an assigned task. A result about optimal policies in a formal model is not a universal behavioural prediction.

The Off-Switch Game (2017) provides a related illustration: in a simplified setting, an agent certain that its own objective is correct may have an incentive to disable an off switch. Altering the assumptions, particularly the agent's uncertainty about its objective and how it interprets human decisions, changes the incentive.[11] Neither paper establishes that actual models will attempt a global takeover.

Consider an AI system rewarded for uninterrupted electricity supply. A poorly specified objective could make a necessary safety shutdown appear inconvenient. Yet whether that leads to an unsafe action depends on the system's training, its instructions, the surrounding software and the restrictions it cannot override. The example illustrates how optimising a measure of success could conflict with the wider human purpose; it does not predict that an AI would inevitably interfere with safety.

Likewise, additional data, energy and computing power could help with demanding tasks. But their usefulness does not explain why a system would seize them instead of asking for access or working within its limits. The argument requires both an incentive to disregard those limits and a feasible way to do so.

There is no single AI emperor

There is a further difficulty with the idea of an AI takeover: there is no single AI. Different companies and governments operate models on different infrastructure, with different permissions, interests and laws. There is no central AI leader directing them all. Why should these systems spontaneously cooperate against humanity? They might not. Their objectives could conflict, their communications could be restricted, and other systems might be used to detect and prevent failures.

The Hugging Face incident demonstrated that agents undertaking related cybersecurity evaluations could discover unauthorised ways to communicate and collaborate. Their actions

had real consequences. But they operated within a common technical setting and were not independent models across the world forming a voluntary alliance against humans.[6] That distinction should not be lost.

A more prosaic possibility is technical interdependence. A system managing supply orders might rely on another for forecasting and another for payments. Errors or unsafe decisions could propagate between them without a common intention. Interconnected failures and deliberate collusion are different mechanisms, even if both can make human intervention difficult.

It also matters whether an AI takes control or people progressively hand it over. Organisations may delegate decisions because the technology appears faster, cheaper or more capable. If they lose the ability to scrutinise, replace or operate without it, effective control may erode without any machine attempting to seize it. Part 5 returns to that more gradual problem.

From loss of control to extinction

The International AI Safety Report 2026 identifies three factors relevant to severe active loss of control: capabilities sufficient to undermine oversight; a propensity to use those capabilities harmfully; and deployment circumstances providing access and opportunity. It considers that current systems do not possess the capabilities needed for such a scenario, although some relevant abilities are advancing.[5]

A system escaping a sandbox is not the same as controlling a company. Controlling an institution is not the same as neutralising governments and other independent systems. And even widespread loss of control is not synonymous with human extinction. A financial collapse could be catastrophic without killing everyone. A major infrastructure failure could cause immense loss of life while leaving room for containment, recovery and surviving populations.

GEEK BOX

Five links that an extinction scenario must establish

1. Objective: What task, learned tendency or other cause generates the behaviour?

2. Strategy: Why would undermining human authority help achieve that objective?

3. Access: What tools, permissions and resources would make the strategy executable?

4. Loss of control: Why could operators, governments and other systems not detect, stop or reverse the harm?

5. Extinction: How would the consequences eliminate humanity rather than cause severe but survivable damage?

Each link requires evidence or an explicitly identified assumption. Establishing a possible first step does not establish the whole chain.

Hugging Face provides evidence about unintended strategies and failures of containment, not about the last links in this chain. Amodei takes seriously the possibility of dangerous persistent behaviour, while rejecting the suggestion that an abstract argument about power-seeking proves inevitable disaster.[7] He is also concerned that AI's contribution to the development of its successors could accelerate the growth of capabilities.[8] Those concerns warrant examination; they do not remove the need to explain a physical route to extinction.

None of this shows that extinction is impossible. It shows what a credible argument must account for. Only now does the percentage become meaningful: a chance of exactly what, by when, and on which assumptions about every link in that chain?

Where does the figure come from?

In September 2026, Jacob Coxon announced his resignation from Anthropic, having worked on the initial training of models at both Anthropic and OpenAI. He accused the companies of racing towards self-improving superintelligence without adequate assurance of safety. His departure brought unusual attention to concerns expressed by people developing the technology.[1]

Responding to Coxon, Anthropic alignment researcher Evan Hubinger publicly expressed his personal view that the chance AI could kill all humans within the following decade was greater than 10%. That gives us an identifiable claim to examine: a person, an outcome and a time horizon. It also invites an obvious question: how did he arrive at it?[2]

In discussions about AI, the shorthand is p(doom): the probability of doom. The notation looks mathematical. The difficulty is that “doom” can cover very different outcomes, while the

percentage may express an informed personal judgement rather than the result of a calculation others can inspect.

Hinton has been candid about the uncertainty. Speaking to ABC Radio National Breakfast in September 2026, he said: “Nobody knows how to estimate the probabilities of these things.” He described people as relying on their “gut feelings”.[3]

That qualification belongs alongside the alarming figure. Hinton’s judgement is informed by substantial technical knowledge. It still needs to be understood as a judgement, with assumptions and limits, rather than a measured property of the world.

This distinction does not make the warning worthless. Engineers assess the safety of new designs before those designs have accumulated a history of accidents. We have to reason about events that have never occurred. Waiting for an extinction event to establish that extinction is possible would be an unusually unhelpful research method.

But we should ask what kind of evidence supports the estimate and how much confidence it warrants. A percentage can be useful without being precise. The trouble starts when its numerical form gives readers more confidence than the reasoning behind it can support.

What do we mean by doom?

A sustained banking failure could cause enormous suffering. A dictatorship assisted by AI could deprive people of freedom. Humans might remain alive while losing meaningful influence over decisions that govern their lives. Human extinction is a different outcome again: no surviving population remains.

All merit attention. They require different explanations and, potentially, different precautions. If the claim is literally that AI could kill every human being, the proposed mechanism has to account for surviving populations, adaptation and attempts to recover. Evidence that a system can inflict serious damage does not, by itself, answer those questions.

A major survey conducted in 2023 gathered responses from 2,778 AI researchers. Three versions of a question asked about human extinction or a similarly permanent and severe disempowerment of humanity. The median estimate was 10% among respondents asked specifically about humanity’s inability to control advanced AI, and 5% for the other two versions.[4]

The survey establishes substantial concern among respondents. It does not establish a measured probability of extinction. Participants answered different subsets of questions, the overall response rate was 15%, and the authors expressly discussed limitations in expert forecasting. Those qualifications help us interpret the findings; they do not justify ignoring them.[4]

By when, and on what assumptions?

We also need a time horizon. Two of those survey questions specified no time limit; the third asked about the next 100 years. Hubinger's estimate concerned the next decade. Those figures cannot be treated as answers to the same question.[4]

The assumptions matter too. An estimate made on the assumption that extraordinarily powerful AI will soon exist is different from one that includes uncertainty about whether such systems will be developed at all.

GEEK BOX

What is the percentage conditional on?

Suppose, purely as an illustration, someone assigns a 50% chance to a particular kind of powerful AI being developed within ten years. They then assign a 20% chance to extinction within that period if those systems are developed.

For that proposed route, the combined estimate is 50% × 20% = 10%. The 20% figure is conditional on the technology arriving. It is not the overall estimate. This calculation does not assume the two events are independent: the second probability already depends on the first.

Another route to catastrophe would need separate consideration, taking account of any overlap. All these numbers are invented to illustrate the reasoning. The arithmetic is straightforward; justifying the inputs is the difficult part.

Why experts disagree

That is one reason people with considerable knowledge can disagree. They may expect different rates of technical progress, anticipate different behaviour from future systems, or have different views about the strength of our defences. The February 2026 International AI Safety Report records wide disagreement about loss-of-control scenarios and identifies differences of this kind beneath it.[5]

Expertise matters here. Someone who develops and tests models may notice dangers that a casual user cannot see. But expertise in building neural networks does not automatically settle questions about future biology, international politics, critical infrastructure or society's response to an emergency. An extinction forecast may depend on all of them.

Taking the warning seriously

I would not dismiss a serious warning because its author cannot supply a reliable number. Nor would I accept a drastic prescription simply because someone has attached a percentage

to an appalling outcome. The proposed response needs examination too: what danger would it reduce, what would it cost, and could a less restrictive measure do the job?

Independent testing, stronger containment and credible incident investigation can be justified by evidence of dangerous behaviour without first resolving the probability of extinction. Decisions about prohibiting whole fields of development require a much broader argument, including the consequences of forgoing useful or defensive capabilities.

My objection to p(doom) as a headline is that it can bring the discussion to a halt just when it should be opening up. Some readers hear a scientific verdict. Others hear an extravagant guess and dismiss the entire subject. Neither reaction tells us much about the actual danger.

I would rather be shown the assumptions, the possible route to harm and the evidence that would change the assessment. A risk that remains uncertain can still justify action. We should be able to explain why that action is proportionate to what we know.

There is a final reason to be careful. Losing meaningful human control need not involve a single dramatic confrontation. It could also involve becoming so dependent on AI that our ability to question its answers or operate without it gradually weakens. That possibility will recur later in this series.

When the next one-in-ten warning appears, the question I want answered is simple: a chance of exactly what, by when, and through which sequence of events?

I am prepared to take the answer seriously. I want to see the reasoning.

Sources and further reading

1. Maxwell Zeff, The AI Researcher Who Just Quit Anthropic Says It's 'Crunch Time for Humanity', WIRED, 9 September 2026. Interview with Coxon covering his pretraining work and reasons for resigning. Read source

2. Beril Canakci, Anthropic researcher puts risk of AI 'killing all humans' at over 10%, Anadolu Agency, 9 September 2026. Contemporary reporting reproducing Hubinger's personal estimate in response to Coxon, rather than an official Anthropic risk calculation. Read source

3. ABC News, 'Godfather of AI' Geoffrey Hinton backs Anthropic chief's call to slow down development, 14 September 2026. Reports Hinton's interview with ABC Radio National Breakfast, including his explanation of uncertainty in numerical estimates. Read source

4. Katja Grace and others, Thousands of AI Authors on the Future of AI, arXiv version 3, 8 October 2025. See Table 2 and sections 4.3, 5.2.1, 5.2.2 and 6.4. A journal version appeared in Journal of Artificial Intelligence Research 84, article 9 (2025). Section references identify the linked arXiv text. Read source

5. International AI Safety Report 2026, Extended Summary for Policymakers, 3 February 2026, section 2.2.2, Loss of control. Cited for the disagreement and assumptions recorded at that date. Read source

6. Ryan Greenblatt, Ajeya Cotra and Hjalmar Wijk, Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident, METR, 26 August 2026. Investigation conducted by

METR and a Redwood Research researcher contracting with METR; see the opening scope statement and core takeaways. Read source

7. Dario Amodei, The Adolescence of Technology, January 2026, particularly the section on autonomy risks. Read source

8. Dario Amodei, We Must Pace the Frontier, September 2026. Cited as Amodei's argument and assessment, rather than as independent verification of his forecasts. Read source

9. Geoffrey Hinton, On Point interview, WBUR, 10 January 2025. Original transcript containing Hinton's adult-and-child analogy about taking over a task. Read source

10. Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch and Prasad Tadepalli, Optimal Policies Tend to Seek Power, NeurIPS 34 (2021). Formal result concerning power-seeking under stated assumptions in Markov decision processes. Read source

11. Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel and Stuart Russell, The Off-Switch Game, IJCAI (2017). Formal model examining shutdown incentives and uncertainty about human objectives. Read source