Could AI Destroy Humanity?

Part 5: What Happens If We Stop Thinking for Ourselves?

Peter Fields · First posted: 5 November 2026

THE 2-MINUTE SUMMARY

The final risk considered in this series is not a machine taking control, but people gradually surrendering the ability to assess decisions made with AI. Generative systems can produce convincing reasoning and useful answers, yet a good result does not necessarily mean the user has developed the knowledge or judgement to verify it.

This article examines evidence about critical thinking and AI-assisted learning, the difficulty of meaningful human oversight, and how widespread dependence can become a shared institutional weakness. It distinguishes an AI system actively undermining control from people gradually surrendering the ability to exercise it. Neither form of loss of control is inevitable, but the two could reinforce one another. What capabilities must people and institutions retain to recognise, challenge and recover from failure?

Imagine a committee approving an AI-written recommendation. The analysis is lucid. The figures look plausible. Someone asks whether the conclusion is sound. Everyone has read the summary, but nobody has examined the underlying evidence. The recommendation passes.

The machine has not seized power. The people in the room have made a decision. Yet how much of the judgement was really theirs?

This final part turns from what AI might do to us towards what we might gradually stop doing for ourselves. My concern is not that every use of AI weakens the mind. It is that we may confuse getting better answers with becoming better able to judge them, and build institutions around that confusion.

We have always borrowed other people's thinking

Books, calculators, colleagues and professional advisers let us achieve things we could not manage alone. Nobody needs to derive every mathematical formula before using a spreadsheet. A society in which everyone must personally verify everything would scarcely function.

Delegation is therefore unavoidable, and often sensible. The useful question is what competence must remain with the person delegating, and what reliable ways of checking the work exist elsewhere. I can rely on an engineer without becoming one. That does not mean accepting a bridge design merely because its explanation sounds confident.

Generative AI makes delegation unusually convenient. The same interface can offer a factual answer, an interpretation of evidence, a proposed strategy and a persuasive explanation of why that strategy is right. The user needs to distinguish those different kinds of claim, even when they arrive in one smooth paragraph.

Saving effort is a benefit. Giving up the ability to recognise when that effort is needed is a different proposition.

What the research supports

A 2025 study by researchers at Microsoft Research and Carnegie Mellon University surveyed 319 knowledge workers, collecting 936 examples of AI use. Greater confidence in AI was associated with less reported critical thinking. Greater confidence in participants' own ability to perform the task was associated with more reported critical thinking. Some also described difficulty checking answers in unfamiliar areas. Participants described a shift towards checking and integrating the output, rather than producing it themselves.[1]

That was a survey of people's perceptions, not a test showing that their intelligence had declined. Self-confidence was not an objective measure of expertise. The authors explicitly say their analysis does not establish causation. Reduced effort might reflect useful assistance, less scrutiny, or difficulty distinguishing the two. It would be wrong to turn this into "AI makes people stupid".

A different study offers firmer evidence within a narrower setting. In a randomised experiment involving nearly 1,000 pupils at one Turkish high school, GPT-4 tools improved maths practice scores by 48% with the basic version and 127% with the guided tutor, relative to the control group. But pupils given the basic version subsequently scored about 17% lower than the control group on exams without assistance. The guided tutor, designed with teacher input to guide learning, produced unaided exam results that were statistically indistinguishable from those of the control group.[2]

Pupils' impressions were also too optimistic. Those using the basic tool did not perceive that they had performed worse or learned less; those using the guided tutor believed their exam performance had improved. Feeling helped was not a reliable measure of what they had learned.[2]

The finding concerns short-term learning in a particular setting. It does not establish permanent damage or the effects of every AI tutor. It does show why the quality of an assisted performance cannot, on its own, tell us what the person has learned.

GEEK BOX

Three different meanings of "getting better"

Suppose someone writes a stronger report with AI. At least three things might have improved: the finished report, their skill at using the tool, or their understanding of the subject. These can improve together, but one does not establish the others.

To assess the report, examine its evidence and reasoning. To assess learning, try a related problem without the assistance. To assess oversight, test whether the person can identify and explain a consequential mistake. Those are different questions, requiring different evidence.

Working unaided is not the only valuable capability. Effective use of AI is a skill too. The point is to measure the capability you need, rather than assume that an impressive output proves all of them.

The paradox of a better system

This problem predates chatbots. In her 1983 paper Ironies of Automation, psychologist Lisanne Bainbridge examined systems that automated normal operations while leaving people responsible for difficult exceptions. Automation could remove the practice through which operators maintained the skills needed to intervene. The most successful systems, rarely requiring human action, might consequently need particularly substantial investment in training.[3]

That was an analysis of industrial automation, not an experiment proving what modern language models do to their users. But the question transfers: are we designing a human role that can actually be performed, or merely leaving a name against the responsibility?

A plausible difficulty with increasingly reliable AI is that confidence in it may be well earned for ordinary tasks. People have reasons to use it. An organisation may then reduce the time, staffing and practice that allowed someone to detect the unusual failure. Its immediate results could improve while its ability to cope with an exception deteriorates. That is a risk to investigate, not an inevitable sequence.

It follows that accuracy and dependence deserve separate attention. Improving the model can reduce mistakes. It does not automatically preserve the expertise needed when a consequential mistake gets through.

A signature is not a safety system

In Part 2, I argued that AI does not need hatred or consciousness to cause harm. Here is a human counterpart: a person does not need to be careless to provide ineffective oversight. They may be conscientious, overworked and unable to obtain the information needed to challenge the recommendation.

Consider an imagined manager expected to approve thirty AI-generated assessments before lunch. Requiring a signature makes the decision traceable. It does not establish that the manager can identify a missing fact, test an assumption or recognise an answer outside the system's competence.

Nor is "ask another AI" a complete solution. A second assessment may help, but agreement alone cannot establish independence or truth. Two tools might rely on the same erroneous source. A generated explanation is something to examine, rather than a substitute for the evidence it purports to explain.

The practical test should be demanding but achievable. For an important decision, can the responsible people establish what evidence supports it, identify a plausible reason it could be wrong, and obtain a check appropriate to the consequences? They need not reproduce every computation inside a model. They do need a defensible basis for acting.

Oversight also requires permission to disagree. If rejecting the recommendation brings delay, criticism or an impossible workload, the organisation should not be surprised when approval becomes the default.

Dependence can become a shared weakness

Individual dependence is only part of the picture. When many institutions rely on the same suppliers, assumptions or methods, a shared weakness can have wider consequences.

The Financial Stability Board's 2024 report identifies dependencies on concentrated AI service providers and the possibility that similar models and data could produce correlated financial decisions. It also recognises benefits and substantial uncertainty about how AI is being used.[4] This is an analysis of vulnerabilities, not a prediction that a chatbot will bring down the banking system.

The distinction matters. Imagine several organisations reaching the same mistaken conclusion through a common model or source. Each could appear to have conducted its own review while sharing the same weakness. The appearance of independent judgement would be misleading. Whether this becomes dangerous depends on what decisions the systems influence and how those decisions interact.

My concern extends beyond finance. If an institution becomes unable to assess its supplier's claims, challenge an output or commission a credible alternative, formal ownership of the decision may conceal considerable practical dependence. That is an inference about how control can be lost through organisational choices, not evidence that every AI user is already in that position.

There is a governance question here as well as a technical one. An organisation should remain answerable for the objectives it pursues and the trade-offs it accepts. A tool may help compare options. "The AI recommended it" cannot explain why those objectives, assumptions or consequences were acceptable.

How does dependence relate to the other risks?

Dependence is not a risk I would confidently rank above the others. A major banking failure can cause immediate, extensive harm. Human extinction is a different order of consequence again. Comparing those risks requires judgements about likelihood, severity and recovery that the evidence on AI dependence cannot settle.

In Parts 1 and 2, we distinguished a system taking unauthorised actions from humans losing control of what those actions could achieve. That distinction has a counterpart here. **Active loss of control** occurs when an AI system evades restrictions or undermines oversight while acting. **Passive loss of control** develops when people or institutions delegate so much that they can no longer scrutinise decisions, intervene competently or operate without the system. The second does not require the AI to seek power at all.

These routes may interact. A technical failure could be easier to detect if experienced operators can challenge the output and recognise an anomaly. The same failure might prove harder to contain if those skills, independent data or recovery procedures have withered. Conversely, a system that appears highly reliable can encourage deeper delegation, making a later failure more consequential. This is a plausible interaction, not evidence that reliance on AI necessarily leads to catastrophic loss of control.

Formal authority should therefore not be confused with practical control. A manager may retain the right to reject an AI recommendation while lacking the information, expertise, time or credible alternative needed to exercise that right. That is why the evidence about learning, institutional dependence and meaningful oversight belongs in the same discussion as technical safety.

But these risks need not compete for attention. Dependence could make a technical failure harder to detect or recover from. It could also weaken the scrutiny of proposals intended to prevent such failures. Our ability to question AI is part of how we manage its other risks.

A loss of competence might accumulate through ordinary decisions. A disruption could then expose it abruptly. That possibility gives us a reason to preserve capability before it is needed, without assuming that all AI harm will arrive slowly or with ample warning.

Equally, a person who cannot complete a task without a particular tool is not necessarily worse off. The tool may give them access to work or knowledge previously beyond reach. The concern becomes sharper where an essential dependency is poorly understood, difficult to replace, and unsupported by effective checks.

What should we keep capable of doing?

I would begin with the decision, rather than a general instruction to use less AI. Which judgements matter enough to require an independent check? Which capabilities are needed to provide it? How would the organisation discover that those capabilities were no longer there?

For individuals, that can mean making an initial assessment before asking for a recommendation, reading the material that could change the answer, or testing whether they can explain the reasoning without the generated text in front of them. These are practical habits to try, not a proven prescription that prevents every form of dependence.

AI can be used within that process. Ask it to challenge an assumption, identify contrary evidence or explain a difficult concept, then examine the result. The safeguard lies in what the user does with the response. Adding “be critical” to a prompt does not ensure that the ensuing criticism is sound.

Institutions have larger responsibilities. People doing consequential checks need access to evidence, protected time and routes to obtain specialist help. Quality assessments should include the ability to recognise exceptions, rather than reward only speed and the appearance of a polished answer. Where continuity matters, recovery arrangements need to be exercised, not merely documented.

Even stopping systems can have consequences. The Financial Stability Board warns that simultaneous shutdowns of automated systems could deepen market stress.[4] Effective control therefore includes planning how to stop, continue safely and recover.

That does not require everyone to work indefinitely without computers. A realistic alternative might involve another independently assessed system, narrower operations or a safe suspension of the affected activity. The right arrangement depends on the service. An alternative that exists only on paper is of little comfort when it is needed.

There is also a longer-term question: who will acquire the experience needed to perform these checks if organisations stop giving beginners meaningful work? That deserves the separate series on trainees and professional development. Here, the immediate point is that retaining today’s experts is not a complete plan for maintaining tomorrow’s competence.

The question this series leaves us with

So could AI destroy humanity? I would not claim that it is impossible. Nor do I think the evidence examined here establishes how likely it is. Specific dangerous capabilities and failures of control deserve attention. They do not, on their own, demonstrate a complete pathway to humanity’s destruction. An expert’s numerical estimate can express a reasoned judgement under uncertainty. It should not be presented as a measured probability.

Across these five articles, I have tried to replace a dramatic headline with questions that can be examined. What objective or behaviour could cause harm? What capability and access would it require? What evidence supports each step towards the feared outcome? Who can intervene, and would they retain the knowledge and authority to do so effectively? Those questions apply to developers, users and regulators alike.

The absence of human emotions in an AI system would not establish safety. Nor does a human origin mean that every harmful action requires a malicious person to order it. Systems can act

unexpectedly after people have trained them, given them objectives and connected them to the world. Our responsibility includes anticipating failures we did not intend.

That is why I would resist both an assurance that intelligence inevitably produces human-like ambition and an assurance that human involvement guarantees control. Capability, access, incentives and the effectiveness of safeguards deserve scrutiny. Whether a machine is conscious does not settle those practical questions.

In the closing chapter of my book, Am I Thinking? The Journal of an AI that Isn't Sure, Alex, the book's AI narrator, turns the question back towards the reader. The concern is whether we can still explain an answer, challenge it and decide when to reject it. I think that is also where this series should end.

We should judge an AI system partly by what it enables people to do, and judge its deployment partly by the capacity it leaves them to question the result. That capacity is worth preserving even when the answers are mostly good.

An off switch can stop a particular system. It cannot restore expertise that an organisation has allowed to disappear. If we intend to remain in control, we need to preserve the ability to decide what happens after we press it.

Sources and further reading

1. Hao-Ping (Hank) Lee and others, The Impact of Generative AI on Critical Thinking, CHI 2025, DOI: 10.1145/3706598.3713778. See pp. 9, 11 – 12 and 14 – 16 for confidence, barriers to checking outputs, discussion and limitations. An English-language self-report survey, not a causal or longitudinal test of skill loss; self-confidence was not an objective measure of expertise.Read source

2. Hamsa Bastani and others, Generative AI Without Guardrails Can Harm Learning: Evidence from High School Mathematics, PNAS 122 (26), e2422633122 (2025). See p. 4, Table 1, Main Results and Student Perception; p. 7 discusses limitations. The 48%, 127% and 17% are relative changes, not percentage points; outcomes were short-term. The 20 August 2025 correction concerns an author's affiliation and does not revise the findings.Read source· Published correction

3. Lisanne Bainbridge, Ironies of Automation, Automatica 19 (6), pp. 775 – 779 (1983), DOI: 10.1016/0005-1098(83)90046-8. See sections 1.1, 2.2 – 2.3 and 3. Analysis of industrial automation, used here to frame questions about oversight, rather than as direct evidence about generative AI.Read source

4. Financial Stability Board, The Financial Stability Implications of Artificial Intelligence, 14 November 2024. See sections 4.1, 4.2.1 and 4.2.2, particularly printed pp. 19 – 22. The warning about simultaneous shutdowns appears on p. 21. Identifies potential shared dependencies and correlated behaviour, alongside benefits, uncertainty and circumstances in which risks may be reduced.Read source

Earlier in this series: Part 1 · Part 2 · Part 3 · Part 4