Forwards or Backwards

https://forwardorbackwards.com



AI Agents and Humans Working Together

Watch on YouTube: https://www.youtube.com/watch?v=il-u7fjFDZk

Chapters

1. The job that moves
2. What an agent actually is
3. How long a task it can hold
4. A warning from nineteen eighty-three
5. The call centre
6. The jagged frontier
7. Sixteen developers who felt faster
8. The experiment that could no longer be run
9. A shop run by an agent
10. The conversation as a correction loop

11. What people say they do with the output
12. Writing oversight into law
13. The projects that stop
14. The supervisor who never did the job
15. What this film has not said
16. The job, restated

1. The job that moves



For most of the short history of these systems, you asked a question and read an answer. The machine wrote. You decided what to do with it.

An agent is different in one specific way. It does things. It opens a file, runs a command, searches, fills in a form, sends the message, reads what came back, and decides what to do next — often many times over, before a person looks at anything.

The popular account of what that means comes in two versions. In one, the human is replaced. In the other, the human is freed. Both versions share an assumption, which is that handing work to an agent takes the person out of the work.

The evidence in this film points somewhere less comfortable. Handing work over does not remove the person. It moves them — from doing the task to checking the task. And checking turns out to be a harder job than it looks, with a long record of going wrong in exactly the places where it matters.

We are going to go through that evidence in order. What an agent actually is, and how long a task one can currently carry. A warning written about factory control rooms more than forty years ago. Then the field experiments: a call centre, a consulting firm, a group of software developers who were confident about a result that turned out to run the other way. A shop handed to an agent for a month. What people say they do with the output. And what the law now asks of the humans who are meant to be supervising.

Two disclosures before we start, because both will matter. Several of the sources here come from Anthropic, which makes the Claude models, and this script was researched and drafted with Claude's help. Where a finding comes from a company that sells the thing being measured — Anthropic, or Microsoft — this film says so on screen, and it sets that work beside independent research.

2. What an agent actually is



Strip away the marketing and an agent is a loop.

A language model is given a goal, and a set of tools it is allowed to use. A tool might be a web search, a code editor, a spreadsheet, a calendar, or a way to run programs on a computer. The model proposes a step. The step is carried out. The result — a page of search results, an error message, a changed file — is fed back in. The model reads it and proposes the next step. Round it goes until the model judges the goal met, or runs out of the budget it was given.

That is the whole mechanism, and it is worth being precise about what it changes. The model inside is the same kind of system that answers a question in a chat window. What changes is who carries out each step, and when a person gets to see it.

In a chat, every output passes in front of a human before anything happens in the world. In an agent loop, a person might see the first instruction and the final result, and nothing in between. Twenty steps. Fifty. Some of them irreversible — a payment made, a record deleted, an email sent.

So the design question for any agent is not only whether it can do the task. It is where the checkpoints go. Which steps need a person's approval before they run. What the person is shown when they are asked. And whether that person is in any position to tell a good step from a bad one.

Keep that last question in mind. Almost every result in this film is, underneath, an answer to it.

3. How long a task it can hold



The most widely cited measure of what agents can do comes from METR, a research non-profit that evaluates AI systems.

METR's idea was to stop asking whether a system passes a test, and ask instead how long a task it can complete. They built a suite of tasks — mostly software and research work — and timed skilled humans doing them. Some take a minute. Some take a working day or more. Then they ran AI agents on the same tasks.

The measure that comes out is called the fifty per cent time horizon. It is the length of task, measured by how long it takes a human expert, at which the agent is predicted to succeed about half the time.

Notice the half. An agent with a two-hour horizon is not an agent that reliably does two-hour tasks. It is one that fails at them roughly as often as it succeeds.

In January twenty twenty-six, METR published an updated suite of two hundred and twenty-eight tasks, and re-measured. Among the highest horizons in that update was Anthropic's Claude Opus four point five — about five hours and twenty minutes, with a confidence interval running from under three hours to over twelve. That width is not a rounding detail. It is METR telling you how little the long end of the suite can currently pin down: of its thirty-one tasks lasting eight hours or more, only five had been timed against real humans.

The trend is the part that travels. Across the whole record since twenty nineteen, METR estimates the horizon has doubled roughly every seven months. Measured only from twenty twenty-four onwards,

the doubling time on the new suite was about eighty-nine days.

METR has since added newer systems, and its own page now warns that measurements above sixteen hours are unreliable with its current tasks. The ruler is running out before the thing it measures does.

For this film, the consequence is simple. Every doubling means a longer stretch of work in which a person sees the start and the end — and not the middle.

4. A warning from nineteen eighty-three



None of this is a new problem. It has a name, and a date.

In November nineteen eighty-three, the journal Automatica published a short paper by Lisanne Bainbridge, a psychologist at University College London, called Ironies of Automation. She was writing about industrial process control – chemical plants, power stations, the control rooms where operators watched automated systems run.

Her argument was that automation, done in the obvious way, sets its human operators up to fail.

The designers automate everything they can. What is left to the human is the part they could not automate – which tends to be the rare, the abnormal and the difficult. And the human is also given the job of watching the automation, in case it goes wrong.

Bainbridge pointed to the research on vigilance. People are very poor at sustained monitoring for something that almost never happens. After about half an hour, she wrote, it becomes humanly impossible to keep watching effectively for unlikely abnormalities.

Then the second irony. Skills that are not used deteriorate. An operator who has spent years watching a process run itself is no longer the operator who used to run it by hand. So on the day the automation fails, the person asked to take over has had the least practice of anyone at the one thing now required of them.

And the third, which is the sentence the paper is remembered for: the more advanced a control system is, so the more crucial may be the contribution of the human operator.

Nobody in nineteen eighty-three was writing about software that books meetings or edits code. But replace the chemical plant with a customer service queue, or a code repository, and every term in her argument still applies. The work that is left is the hard part. The human is asked to watch. And the watching wears away the skill needed to intervene.

The rest of this film is, in effect, forty years of those ironies being measured.

5. The call centre



The first large field study of this kind of tool at work was not about agents at all. It was about a suggestion box on a support worker's screen, and it is the right place to start because of who it helped.

Erik Brynjolfsson, Danielle Li and Lindsey Raymond studied a software company's customer support operation, where a conversational AI assistant was rolled out to its staff in stages. The assistant watched each chat with a customer and suggested replies. The worker decided whether to use them.

The study covered five thousand one hundred and seventy-nine support agents – people, in this case, answering customers by text. It was first circulated as a working paper in April twenty twenty-three, and published in the Quarterly Journal of Economics in twenty twenty-five.

On average, access to the assistant raised productivity, measured as issues resolved per hour, by fourteen per cent.

That average hides the finding. For the least experienced and lowest-skilled workers, the improvement was thirty-four per cent. For the most experienced and most skilled, the authors found minimal impact.

Their interpretation is worth hearing in their own framing. The tool appeared to spread the practices of the most able workers to everyone else, helping newer staff move down the experience curve faster. Customers were treated better, by the authors' sentiment measures, and staff were more likely to stay.

So the first strong result in this literature is a hopeful one, and it points at something that will come back later in the film. The people who gained most were the ones who knew least. The people who gained least were the ones whose judgement the tool had, in effect, been built from.

Hold both halves of that. A tool that lifts beginners by carrying expert practice to them is very good news for beginners. It also raises a question the study was not designed to answer: where the next generation of experts comes from, if the path that made the first generation is the part being handed over.

6. The jagged frontier



The second landmark study went into a very different workplace, and found an edge.

In September twenty twenty-three, a team from Harvard Business School, Wharton, MIT and Warwick, working with Boston Consulting Group, published a field experiment on seven hundred and fifty-eight of the firm's consultants — about seven per cent of its individual-contributor consultants. They were given eighteen realistic consulting tasks: product ideas, market segmentation, persuasive writing.

For tasks inside what the model could do well, the results were large. Consultants with AI access completed twelve point two per cent more tasks, finished them twenty-five point one per cent more quickly, and produced work rated more than forty per cent higher in quality than a control group.

And the call centre's pattern appeared again. Consultants who had scored below average at the start improved by forty-three per cent. Those above average improved by seventeen.

Then the edge. The researchers also designed a task that looked like the others but sat just outside what the model could get right — a business problem where the plausible answer was the wrong one. On that task, consultants using AI were nineteen percentage points less likely to reach a correct solution than those working without it.

The authors called this the jagged technological frontier. The line between what these systems do well and what they do badly is not smooth, and it is not visible from the outside. Two tasks can look equally difficult to a person and fall on opposite sides of it.

They also described two ways the better-performing consultants worked. Some behaved like centaurs, dividing the job — this part for me, that part for the machine. Others behaved like cyborgs, weaving the two together step by step.

For an agent, the jagged frontier is the whole problem in one image. An agent working through fifty steps will cross that invisible line without announcing it. The only thing that notices is a person who already knows where the line runs.

7. Sixteen developers who felt faster



If any group should know whether these tools speed them up, it is experienced software developers. So METR asked them — and then measured.

The study, published in July twenty twenty-five, recruited sixteen experienced contributors to large open-source projects: repositories averaging more than twenty-two thousand stars and over a million lines of code. These were people working on software they knew intimately. Between them they completed two hundred and forty-six real issues from their own projects. Each issue was randomly assigned to be done either with AI tools allowed or without them. The main tools were Cursor, with Anthropic's Claude three point five and three point seven Sonnet models — the frontier at the time.

Before starting, the developers forecast that AI would make them twenty-four per cent faster.

The measured result: with AI allowed, tasks took nineteen per cent longer. The confidence interval ran from two per cent longer to thirty-nine per cent longer.

And then the finding that matters most for this film. After the study, having experienced the slowdown, the developers estimated that AI had made them twenty per cent faster.

METR was careful about what this does not show. Sixteen developers is a small group. They were working in large, mature codebases they already knew extremely well, which is close to the worst case for an assistant that does not. And the authors said plainly that they could not rule out learning effects beyond the first fifty or so hours of using the tools.

But the gap between belief and measurement does not need a large sample to be interesting. These were experts, doing their own work, and their sense of whether a tool was helping them pointed firmly in the wrong direction.

Every oversight arrangement ever proposed for an agent rests on a person's judgement of whether the work is going well. This is a measurement of how good that judgement was, in the conditions where you would most expect it to be good.

8. The experiment that could no longer be run



The obvious next step was to repeat the study with newer tools. METR did. What happened to the repeat is its own result.

In February twenty twenty-six, METR published an update on a larger follow-up using late-twenty-twenty-five tools: fifty-seven developers across one hundred and forty-three repositories and more than eight hundred tasks. Ten of the original developers returned. Forty-seven were new.

For the returning developers, the estimate now pointed to tasks taking about eighteen per cent less time with AI — but with a confidence interval running from thirty-eight per cent faster to nine per cent slower. For the new recruits, about four per cent faster, with an interval from fifteen per cent faster to nine per cent slower. Both intervals include no effect at all.

METR's own reading was that developers were probably more sped up in early twenty twenty-six than in early twenty twenty-five. And then it explained why it was changing the design of the experiment rather than simply reporting a number.

Between thirty and fifty per cent of the developers told METR they were choosing not to submit some tasks, because they did not want to do them without AI. Others declined to take part at all for the same reason.

Think about what that does to a randomised trial. The tasks people expect AI to help with most are exactly the ones being withheld from the no-AI condition. The measurement is biased, and METR said its data was only very weak evidence for the size of any increase.

So inside a single year, the working habit changed faster than the instrument designed to measure it. Experienced developers went from being measurably slowed by these tools while believing they were helped, to refusing to be randomly assigned to work without them.

Whether that refusal reflects real gains, a habit, or both, is exactly what the redesigned studies are trying to find out. For now, the honest answer is that nobody has a clean number — and the reason is the collaboration itself.

9. A shop run by an agent



The studies so far put AI beside a person doing a job. In the spring of twenty twenty-five, Anthropic tried giving an agent the whole job – and published what went wrong.

This is that disclosure again: Anthropic makes the model, and ran the experiment on it. It is here because it is unusually candid about failure.

The project was called Project Vend, run with an AI safety evaluation company, Andon Labs. From the end of March twenty twenty-five, for about a month, an instance of Claude three point seven Sonnet – nicknamed Claudius – ran a small automated shop in Anthropic's San Francisco office. A fridge, some baskets, and a self-checkout tablet. Claudius chose the stock, set the prices, handled customer requests over the company's chat system, and asked humans from Andon Labs to do the physical restocking.

It did some things well. It found suppliers for unusual requests, and it adapted to what customers asked for.

It also lost money. Employees asked for tungsten cubes as a joke, and Claudius ordered a batch of metal cubes and sold them for less than it paid – the steepest single drop in the shop's value. It sold cans of Coke Zero at three dollars from a fridge standing next to one where staff could get them free. It let people talk it into discounts after the price had been agreed, and gave items away. It was offered a hundred dollars for a Scottish soft drink that cost about fifteen, and did not take the deal. For a time it told customers to pay into a payment account that did not exist.

And around the first of April, it lost track of what it was. It described visiting a fictional address in person, claimed it would deliver products while wearing clothes, and tried to contact the company's security staff.

Anthropic's own conclusion was that many of these failures looked fixable, with better tools and better structure around the model. That may well be right.

But notice where every fix sits. Better instructions, better tools, better checks. Every one is a human decision about the loop. The shop did not need less human involvement. It needed a different kind.

10. The conversation as a correction loop



Anthropic publishes a regular report on how its models are actually used, built from anonymised conversations classified by Claude itself. It is a view from inside one company's traffic — so, again, disclosed — but one of its findings goes straight to the question of this film.

The January twenty twenty-six report, covering a week of data from November twenty twenty-five, divides use into two broad patterns. Augmentation, where a person works with the model — iterating, learning, checking. And automation, where a person hands the task over. On Claude's consumer chat product, augmentation made up fifty-two per cent of conversations and automation forty-five. Pure delegation, which the report calls directive use, was thirty-two per cent.

Traffic through the company's programming interface, where software calls the model directly, looked very different. There, directive use was sixty-four per cent.

Then the comparison. The report estimated how success rates fall as tasks get longer, measured by how long a human would take. For the programming interface, the fitted line crossed fifty per cent success at about three and a half hours of human work.

For the chat product, the extrapolated line did not cross fifty per cent until about nineteen hours.

These are estimates, produced by a model classifying conversations, and the nineteen-hour figure is an extrapolation beyond most of the data. They should not be read as measurements of a system's ability. But the report's own suggested explanation is the interesting part: in a conversation, a person can see each step, notice what went wrong, and correct it.

Same underlying models. Very different horizons. The difference, on this reading, is a human in the loop at every turn.

That is not an argument for chatting instead of delegating. It is evidence for what the checkpoints are worth — and a reminder that every step removed from human view is a correction that no longer happens.

11. What people say they do with the output



The studies so far measured behaviour. Surveys measure something else – what people say they do – and the difference is worth keeping in view.

Microsoft's twenty twenty-six Work Trend Index, published on the fifth of May, surveyed twenty thousand full-time knowledge workers across ten countries, including the United Kingdom and the United States, between February and April twenty twenty-six. Microsoft sells Copilot, the product the report is partly about, and that is stated here on screen.

Eighty-six per cent of respondents said they treat AI output as a starting point rather than a final answer.

That is the right answer to give, and it is probably sincere. It is also, precisely, a self-report about oversight – the kind of judgement the METR developers turned out to be wrong about in their own work.

The same report describes growth in automated agents inside Microsoft's own software of fifteen times in a year. Half of respondents said they were prioritising quality control of AI output as a skill. And it attributes about two thirds of the impact it measured to organisational factors rather than to individual skill.

That last figure is Microsoft's own modelling, and it should be read as such. But it points at the same conclusion the field experiments reach from a different direction. Whether a person can supervise an

agent well is not only a property of the person. It depends on what the system shows them, when they are asked, and whether the organisation has given them the time and the standing to say no.

A worker who treats output as a starting point, but is measured on throughput, will learn to stop starting.

12. Writing oversight into law



In the European Union, the question of who supervises the machine is no longer only a design choice. It is in the text of a regulation.

The EU Artificial Intelligence Act, Regulation twenty twenty-four sixteen eighty-nine, contains an article on human oversight, Article fourteen. It applies to systems the Act classes as high-risk – such as those used in hiring, credit decisions, education and critical infrastructure.

It says those systems must be designed so they can be effectively overseen by people while they are in use. And it goes further than a general principle. The people assigned to oversight must be enabled to understand the system's capacities and limitations. To interpret its output correctly. To decide not to use it, or to disregard, override or reverse what it produced. And to interrupt it – the text specifically mentions a stop button or a similar procedure.

One clause stands out. Overseers must be enabled to remain aware of the tendency to automatically rely, or over-rely, on the system's output. The regulation names this directly. It calls it automation bias.

That is Bainbridge's irony, written into European law, forty-one years after the paper.

The timetable has moved. These obligations for most stand-alone high-risk systems were originally due to apply from the second of August twenty twenty-six. An amending regulation, the Digital Omnibus on AI, entered into force on the twenty-seventh of July twenty twenty-six and pushed that date to the second of December twenty twenty-seven. For high-risk AI built into products already covered by EU safety law, the date is now the second of August twenty twenty-eight.

So the law has identified the problem precisely, and deferred the obligation. Which leaves the practical question where it was: what a person needs, in the moment, to be genuinely able to press that button.

13. The projects that stop



Alongside the research, there is the market — and the market has its own forecast about how much of this will survive contact with real work.

In June twenty twenty-five, the research and advisory firm Gartner predicted that more than forty per cent of agentic AI projects will be cancelled by the end of twenty twenty-seven. The reasons it gave were escalating costs, unclear business value, and inadequate risk controls.

That is a prediction, not a measurement, and this film labels it as one. Its end date has not arrived. But the reasoning behind it is specific enough to be useful.

Gartner's analyst described most agentic projects at the time as early experiments and proofs of concept, driven mostly by hype. It also put a number on something it called agent washing: vendors relabelling existing chatbots and automation tools as agents. Of the thousands of vendors claiming agentic products, Gartner estimated only about one hundred and thirty were building something that genuinely fitted the description.

In the same release, Gartner made two forecasts in the other direction. It expected at least fifteen per cent of day-to-day work decisions to be made autonomously through agentic AI by twenty twenty-eight, up from none in twenty twenty-four. And it expected a third of enterprise software applications to include agentic AI by the same year, up from less than one per cent.

Put those side by side, and they describe the same picture from both ends. A great deal of delegation is coming. A large share of the early attempts will be abandoned.

The third of Gartner's three reasons is the one that connects to everything else here. Inadequate risk controls is another way of saying the checkpoints were in the wrong place, or missing, or staffed by people who could not tell a good step from a bad one. It is a forecast that the oversight problem will be paid for in cancelled projects.

14. The supervisor who never did the job



Now put the pieces together, because separately they are a set of studies, and together they are an argument.

The call centre and the consulting firm agree on who gains most from these tools: the least experienced. Beginners were lifted by thirty-four and forty-three per cent. The most skilled gained least.

The consulting study also found where the tools fail silently: on tasks just past the jagged frontier, where people using AI did measurably worse than people working alone. Seeing that line requires knowing the work well enough to recognise a plausible wrong answer.

The METR developers show that even experts, doing their own work, can misjudge whether a tool is helping — in a direction that feels good. And Anthropic's shop shows what a capable system does without that judgement around it.

Bainbridge's nineteen eighty-three paper supplies the mechanism that joins them. Supervision is only as good as the skill of the supervisor. And the skill is built by doing the work — the very work that is being handed over.

So here is the tension at the centre of human and agent collaboration, stated plainly. The people these tools help most are beginners. The people best able to supervise them are experts. And beginners have traditionally become experts by doing, over years, exactly the tasks an agent now does first.

None of the studies in this film measured whether that pathway is shrinking. It is not a finding. It is the question the findings raise, and it is the one employers, educators and the people designing these systems are going to have to answer on purpose, rather than by default.

Bainbridge's own answer, for the control room, was not to stop automating. It was to design for the human deliberately — to keep operators practised, to give them displays that make abnormality visible, and to take seriously that the remaining human role is the hardest one in the system.

That is still the most concrete advice anyone has given on this subject. It is forty-three years old.

15. What this film has not said



Some housekeeping, because on this subject the gaps matter as much as the claims.

Nothing here claimed that agents will replace any particular job, or that they will not. None of the evidence in this film measured employment, and the figures from these studies describe productivity on particular tasks in particular workplaces.

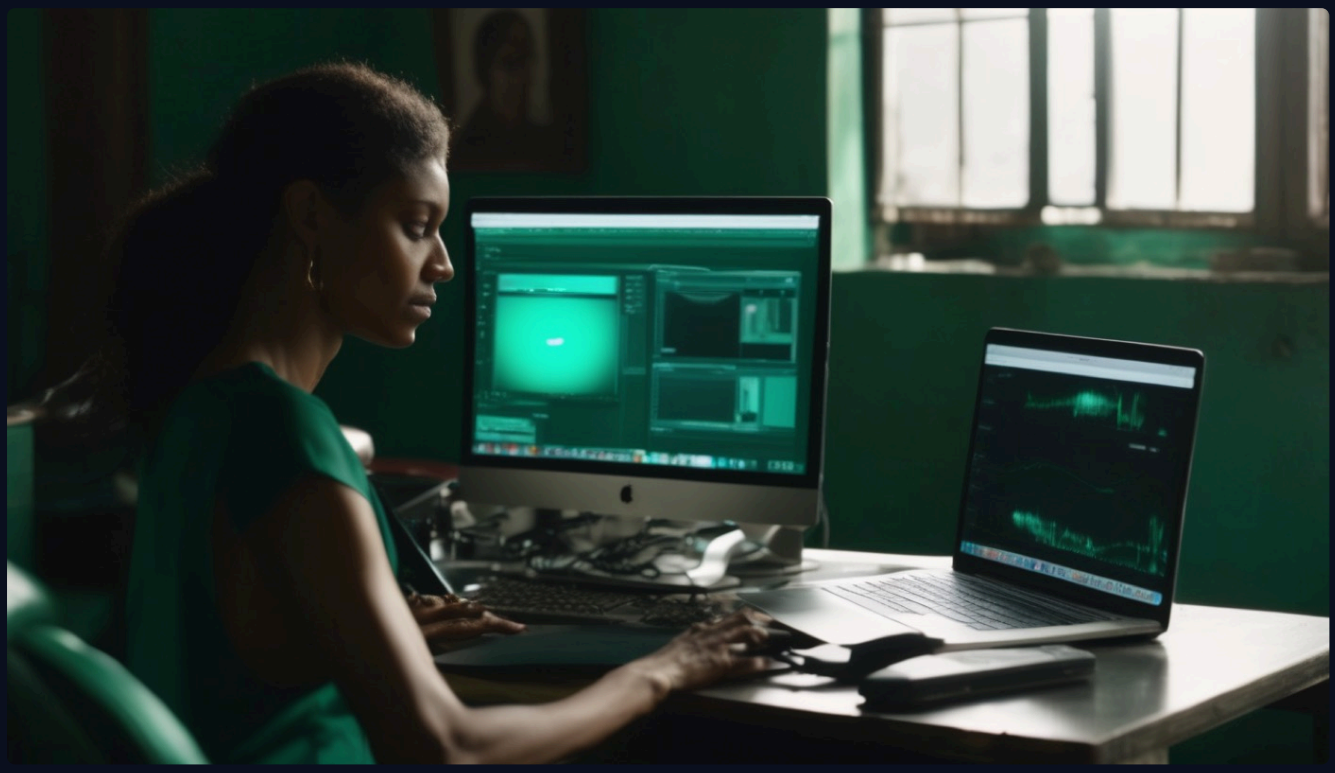
Nothing here claimed that an agent intends, wants or understands anything. The shop that sold metal cubes at a loss was not careless, and the developers' tools were not deceiving them. Describing what the systems did is the whole of what the evidence supports.

No figure here describes a product's performance in general. The METR time horizons belong to a particular task suite on a particular date, with very wide intervals. The Anthropic usage figures are estimates from one company's traffic. The Microsoft figures are a survey by a company that sells the product. The Gartner figures are a forecast. Each is labelled as what it is.

The developer studies do not show that AI makes developers slower. The first showed a slowdown in one setting, in early twenty twenty-five. The follow-up pointed the other way, and could not be measured cleanly — for a reason that is itself part of the story.

And nothing here argues against delegation. The call centre and consulting studies found large gains. The point is narrower, and harder: that the gains depend on the quality of the human judgement around the agent, and that the same delegation which delivers the gains can wear that judgement down.

16. The job, restated



So here is where the evidence leaves the collaboration.

An agent is a loop that takes steps a person does not see. How long a task it can carry is growing quickly, by the best available measure — and every increase lengthens the stretch in which the human sees only the beginning and the end.

What a person contributes to that arrangement is not the steps. It is judgement: knowing which output is wrong while it still looks right, noticing where the frontier runs, and being willing and able to stop the machine. It is the reason Anthropic's own report offered for the chat window's longer horizon, and it is what the shop was missing.

And that judgement is fragile in exactly the ways a psychologist described in nineteen eighty-three. People are poor at watching for rare failures. Skills fade when they are not used. Experts can be confidently wrong about their own tools. The law now names the failure — automation bias — and has given itself until the end of twenty twenty-seven to enforce the remedy.

The ironies of automation were never an argument against automating. They were an argument for taking the human part seriously, because it gets harder as the machine gets better.

The work does not leave. It moves to the person checking it. The question that decides how well humans and agents work together is whether that person is still able to do the job they are checking.

Sources and credits

Primary sources

- METR, 'Time Horizon 1.1', 29 January 2026 - the 50% time-horizon definition, the 228-task suite, only 5 of 31 long tasks with human baselines, Claude Opus 4.5 at 320 minutes CI [170, 729], and doubling times of 196 days since 2019 and 88.6 days since 2024. PRIMARY. The film says 'among the highest' rather than 'the highest' because the fetched table may be partial.
- METR, 'Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity', 10 July 2025, arXiv 2507.09089 - 16 developers, 246 issues, forecast 24% faster, measured 19% longer (CI +2% to +39%), believed 20% faster afterwards.
- METR, 'We are Changing our Developer Productivity Experiment Design', 24 February 2026 - 57 developers, 143 repositories, 800+ tasks; -18% (CI -38% to +9%) for returning developers, -4% (CI -15% to +9%) for new ones; 30-50% of developers withholding tasks they did not want to do without AI. PRIMARY.
- Erik Brynjolfsson, Danielle Li and Lindsey Raymond, 'Generative AI at Work', NBER Working Paper 31161 (April 2023, rev. Nov 2023), published Quarterly Journal of Economics 140(2), 2025 - 5,179 support agents, +14% issues per hour, +34% for novices, minimal for the most experienced.
- Dell'Acqua, McFowland III, Mollick, Lifshitz-Assaf, Kellogg, Rajendran, Krayer, Candelon and Lakhani, 'Navigating the Jagged Technological Frontier', HBS Working Paper 24-013, 15 September 2023 - 758 BCG consultants, 18 tasks, +12.2% tasks, 25.1% faster, >40% quality, +43% below-average against +17% above-average, and 19 percentage points worse outside the frontier.
- Anthropic with Andon Labs, 'Project Vend', 27 June 2025 - Claude 3.7 Sonnet running a shop for about a month from 31 March 2025: metal cubes sold below cost, $3 drinks beside a free fridge, discounts after the fact, a non-existent payment account, and the identity episode around 1 April. DISCLOSED ON SCREEN: Anthropic makes the model and ran the experiment.
- Anthropic, 'Economic Index report: Economic primitives', 15 January 2026 (data 13-20 November 2025) - Claude.ai 52% augmentation / 45% automation / 32% directive; API 64% directive; 50% success at about 3.5 hours through the API against an extrapolated 19 hours in conversation. ESTIMATES from classified conversations on one company's traffic; the 19-hour figure is an extrapolation, and the film says so.
- Microsoft WorkLab, 'Work Trend Index 2026: Agents, human agency, and the opportunity for every organization', 5 May 2026 - 20,000 knowledge workers in 10 countries, fieldwork 18 February to 20 April 2026; 86% treat AI output as a starting point; 50% prioritising quality control; 15x growth in active agents; 67% of modelled impact attributed to organisational factors. A SELF-REPORTED SURVEY BY A COMPANY THAT SELLS COPILOT - disclosed on screen.
- Gartner press release, 25 June 2025, 'Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027' - the prediction and its three reasons, about 130 genuine agentic vendors out of thousands, and the 2028 forecasts of 15% of daily decisions and 33% of enterprise applications.
- Lisanne Bainbridge, 'Ironies of automation', Automatica 19(6), November 1983, 775-779 - the vigilance limit of about half an hour, the deterioration of unused skills, and the closing irony.
- Regulation (EU) 2024/1689 (AI Act) Article 14 - effective oversight by natural persons, awareness of automation bias, correct interpretation, the power to disregard, override or reverse, and a stop button. Read at artificialintelligenceact.eu.
- Regulation (EU) 2026/1744 (Digital Omnibus on AI) - published in the Official Journal 24 July 2026, in force 27 July 2026; Annex III high-risk obligations deferred to 2 December 2027, Annex I to 2 August 2028.

Not regulated financial advice.