



The Competence Paradox

When AI Makes the Team Better and the Expert Worse

About this paper

This paper is for boards and the people who advise them. It addresses one question that productivity figures cannot answer, which is what happens to the judgement of skilled people when a capable system does much of their thinking for them, month after month. The evidence it draws on comes mainly from human factors research in aviation, medicine and decision support, and from the first controlled studies of generative tools in knowledge work.

It is the second of six Ecaveo working papers on the higher-order issues in AI adoption. The strand holds a single argument, that AI adoption is a changing sociotechnical configuration whose governance must keep pace with it. This paper applies that argument to expertise. The people who oversee an AI system are part of the configuration, and their competence changes as the system changes their work, so any governance that treats human oversight as a fixed control is governing something that no longer exists.

Editorial balance. Roughly half of what follows reports published evidence, about a third is my own analysis of what that evidence implies for organisations adopting generative AI, and the remainder, concentrated in the final two chapters, sets out a framework of practices and measures that I propose and that has not been tested as a whole.

HOW TO READ THE EVIDENCE TAGS

Every substantive claim carries a tag stating what kind of evidence supports it. Two tags are combined where a claim rests on both.

PEER-REVIEWED A journal article or peer-reviewed conference paper.

PREPRINT A working paper, preprint or technical report not yet peer reviewed.

STANDARD A published standard or framework, such as ISO or NIST.

POLICY A government, regulator or court document or action.

INDUSTRY A company announcement, press report, analyst or consultancy report.

SCHOLARLY A scholarly book, edited volume or law journal article.

PRACTICE Practitioner observation, including the author's own.

EVIDENCE GAP The available evidence does not answer the question.

ECAVEO The author's own framework or analysis, proposed rather than proven.

CONTENTS

Contents

FRONT MATTER

A note from the author	4
Executive summary	5

PART ONE · THE PROBLEM AND THE EXISTING EVIDENCE

1	The analyst who stopped checking	8
2	Trust is the wrong target	10
3	What automation research already shows	12

PART TWO · THE LONG RUN

4	The long-run problem of skill erosion	16
5	Generative AI raises the stakes	19

PART THREE · GOVERNING FOR RETAINED EXPERTISE

6	Designing for retained expertise	22
7	Measures a board should ask for	24

END MATTER

References	26
About the author	27

ILLUSTRATIVE MATERIAL

Every organisation, person and scenario described as illustrative in this paper is invented. Published research and official documents are cited and numbered, and invented material is never presented as evidence.

Research closed 17 February 2023.
Sources and events are as they stood at that date.

CITATION

Forrest, P. (2023) *The Competence Paradox: When AI Makes the Team Better and the Expert Worse*. Ecaveo Working Papers, AI Adoption: The Higher-Order Issues, Paper 2. Ecaveo, 21 February 2023. Available at: ecaveo.com.

Comments and corrections are welcome and will be acknowledged in later versions.

PAPERS IN THIS STRAND

- Paper one. *Knowledge With an Expiry Date*. Evidence, Evaluation and Decision-Making When the AI System Keeps Changing. 14 December 2022.
- **Paper two. *The Competence Paradox. When AI Makes the Team Better and the Expert Worse*. 21 February 2023.**
- Paper three. *The Unit of Adoption*. Why Organisations Adopt Configurations, Not Models. 31 March 2023.
- Paper four. *Everybody's System, Nobody's Risk*. Accountability Gaps in Distributed AI Governance. 28 April 2023.
- Paper five. *Approved Once, Changed Daily*. From Point-in-Time Ethics to Continuous AI Assurance. 31 May 2023.
- Paper six. *The Cost of Waiting*. Precaution, Opportunity and the Ethics of Not Adopting AI. 29 June 2023.

A NOTE FROM THE AUTHOR

Why I started counting the checks nobody makes

My interest in this problem began with something small. In advisory work I have watched capable teams adopt tools that made them faster, and I noticed that the conversations about those tools were almost entirely about speed and output. Nobody asked how often the people using them still checked the work, or whether they could still do it without help. I asked a few teams, informally, and the honest answer was usually that nobody knew.

Coaching taught me a related lesson years ago. An athlete who trains only with assistance, whether that is a pacer, a machine or a coach calling every split, can look better in training than they are, and the gap shows on the day the assistance is missing. Good coaches build in sessions without the support on purpose, even though those sessions look worse on paper. I have come to think that knowledge work needs the same discipline, and that almost no organisation has it.

Conversational AI tools have made the question urgent. These tools are fluent, fast and often right, which is the profile that the older research on automation says will erode vigilance. My view is that organisations should keep adopting, while measuring something they have so far ignored and protecting a capability that most of them assume will simply persist.

I should be plain about the limits of what I offer. The research base on long-run skill loss is thin outside aviation and medicine, and there is almost nothing yet on generative AI specifically. The practices in the final chapters are my proposals, drawn from the evidence and from practice, and I would welcome challenge to any of them from people who have tried something similar.

Paul Forrest
Founder, Ecaveo
21 February 2023

TWO PAGES THAT CIRCULATE ON THEIR OWN

Executive summary

AI tools can raise the quality and speed of a team's output while quietly weakening the expertise its people need to notice when the tool is wrong, so boards should govern for appropriate reliance and retained skill as well as for throughput.

The research on automation, decades deep, shows that people over-rely on capable systems, that experts are not immune, and that skills left unpractised decay. Generative AI extends that pattern into open-ended professional work, where errors are fluent and harder to spot. Productivity studies, which measure speed on a single task, cannot see the problem. What follows sets out the evidence, explains why the usual remedies fall short, and proposes practices and measures that let an organisation keep the benefit without losing the people who can catch the failure.

1

The right target for AI-assisted work is appropriate reliance, meaning that people accept good advice and reject bad advice.

Trust is an attitude and reliance is a behaviour, and the two can diverge.¹ **PEER-REVIEWED** Governing for more trust, or for more use, says nothing about whether decisions are improving.

2

Over-reliance is persistent, affects experts as well as novices, and is not removed by training or instruction alone.

A review of complacency and automation bias found both effects in naive and expert participants, and found that automation bias could not be prevented by training or instructions.² **PEER-REVIEWED** Physicians' diagnostic accuracy fell when they received inaccurate advice, whatever its purported source.³

PEER-REVIEWED

3

Explanations do not reliably stop people following wrong advice.

In one set of studies, explanations increased the chance that people accepted an AI recommendation whether or not it was correct.⁴ **PEER-REVIEWED** Later work suggests explanations can help when they lower the cost of checking.⁵ **PREPRINT**

4

Skills that automation takes over can decay without anyone noticing until the automation is gone.

Airline pilots retained manual control skills but showed more frequent problems with the cognitive tasks of manual flight.⁶ **PEER-REVIEWED** In one firm, deskilling among accountants surfaced only when an automated system was withdrawn.⁷ **PEER-REVIEWED**

5

Productivity evidence for generative AI is real but measures the wrong thing for this risk.

A controlled study found developers with an AI pair programmer completed a task 55.8% faster than those without.⁸ **PREPRINT** That result says nothing about whether those developers can still detect a flawed suggestion, and no study I know of has yet measured this for generative tools. **EVIDENCE GAP**

6

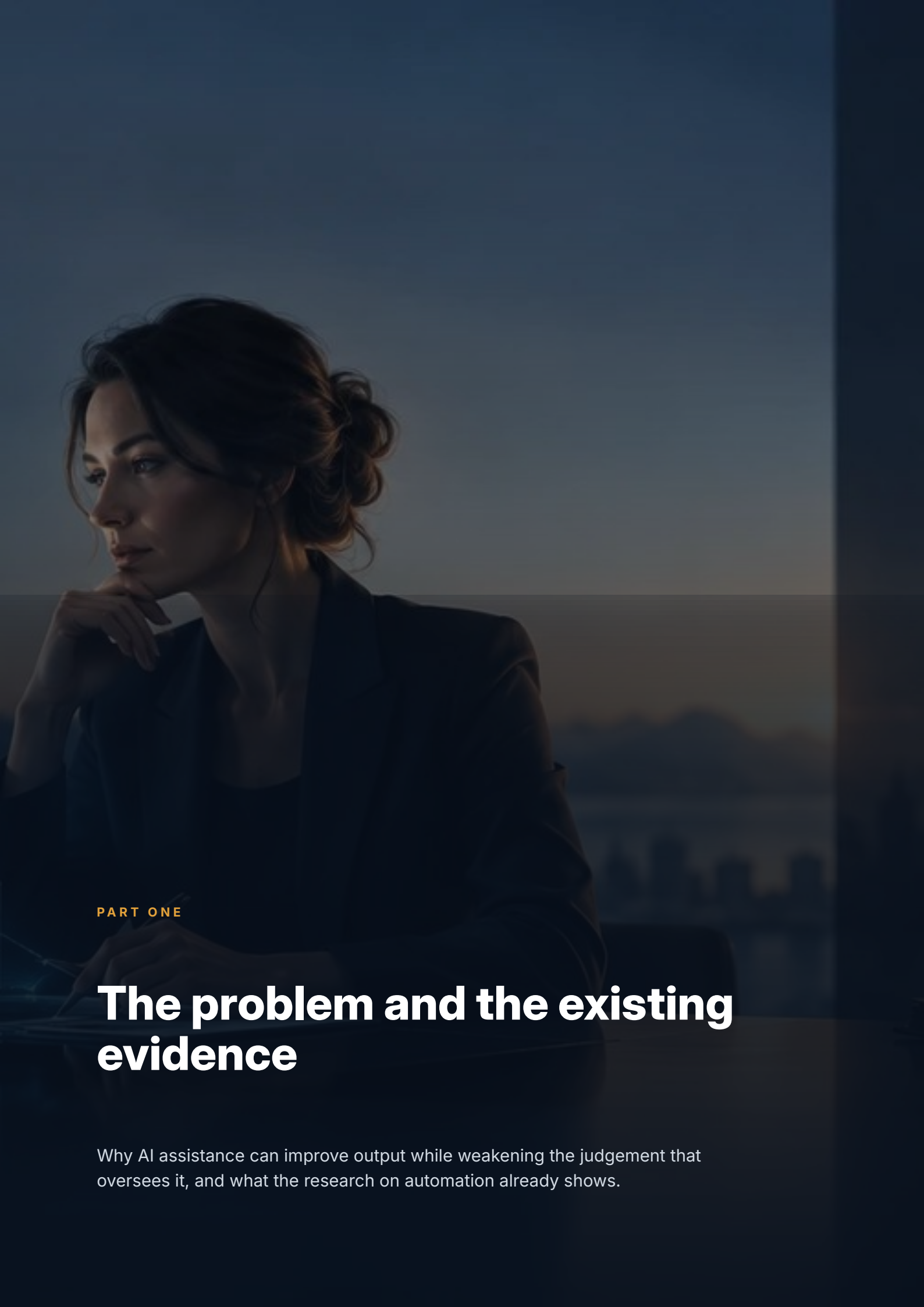
Organisations can protect expertise by designing for it and measuring it directly.

I propose four practices (cognitive forcing, periodic unaided work, disagreement logging and tool-off drills) and a small set of board measures that sit beside productivity. **ECAVEO** The US AI Risk Management Framework itself suggests that data on how often and why people overrule AI output may be useful to collect.⁹ **STANDARD**

What follows for a board

A board should stop asking whether its people trust the AI tools they use and start asking whether they still catch the tools' mistakes. That question needs evidence, and the evidence has to be collected on purpose, because the ordinary dashboards of adoption and output will show improvement right up to the point of failure.

In practice this means three commitments. The organisation should name the roles where an undetected AI error would be costly, require those roles to keep a measured share of unaided work, and report override and error-catch rates alongside productivity at the same cadence. None of this is expensive, although some of it will look like waste to a manager judged on throughput, and the board may need to protect it for that reason.



PART ONE

The problem and the existing evidence

Why AI assistance can improve output while weakening the judgement that oversees it, and what the research on automation already shows.

CHAPTER ONE

The analyst who stopped checking

An organisation can see the quality of its work rise while the people producing it become less able to tell when the machine is wrong. This is the competence paradox, and no productivity dashboard will show it.

The composite below is invented. It draws on patterns that recur in the research literature and in practice, and it is offered to make the mechanism concrete before the evidence is examined.

ILLUSTRATIVE SCENARIO**The quarterly review that nobody questioned**

A senior analyst at a mid-sized financial services firm begins using a conversational AI tool to draft the commentary for a quarterly portfolio review. In the first month she checks every figure and rewrites most paragraphs. By the third month the drafts are good enough that she checks only the numbers that look odd. By the sixth month she reads the drafts for tone and signs them off, because the tool has not made a material error that she can remember, and her manager has doubled the number of portfolios she covers. In the eighth month a draft attributes a fall in one holding to a sector-wide cause that did not happen. The explanation is plausible, well written and wrong. She does not notice, and neither does the reviewer, who has also begun reading for tone. A client does.

Nothing in that story requires negligence. Each step was a reasonable response to a run of good drafts and a growing workload, and the organisation rewarded every step. The analyst's output rose in volume and, on most days, in quality, while her habit of checking faded, and with it, probably, some of the detailed knowledge of the portfolios that checking had kept fresh. The firm measured the first change and had no measure for the second.

1.1 The paradox stated

My central claim is simple to state and uncomfortable to act on. The capability that makes AI assistance safe is the ability to use good advice and catch bad advice, which the human factors literature calls appropriate reliance.¹ **PEER-REVIEWED** That capability depends on expertise, since a person can reject a wrong answer only if they know enough to see that it is wrong. Routine reliance on a capable system, however, removes the practice that keeps expertise alive. The better the tool performs on ordinary days, the less the expert exercises the judgement needed for the day it fails.

ECAVEO

The paradox has two faces. At team level, assistance can lift the average, and early evidence suggests it may lift the less experienced most.⁸ **PREPRINT** At the level of the individual expert, the same assistance may slowly hollow out the skill that made that

person worth employing as a check on the system. The team gets better and the expert gets worse, and the organisation experiences only the first of those until a failure reveals the second.

1.2 Why the question is pressing now

Three months ago the question was largely academic outside a few specialist fields. OpenAI released ChatGPT on 30 November 2022, and in doing so stated plainly that the system sometimes writes answers that sound plausible and are incorrect or nonsensical.¹⁰ **INDUSTRY** On 7 February 2023 Microsoft announced a new Bing that combines search with a chat interface, available that day in a limited preview.¹¹ **INDUSTRY** Google announced its own conversational system, Bard, the day before, and on 8 February Reuters reported that promotional material for Bard contained a factual error about the James Webb Space Telescope, and that Alphabet lost about \$100bn in market value that day.¹² **INDUSTRY**

The Bard episode is useful here for one narrow reason. The error sat in material that a large and well-resourced company had prepared and, presumably, reviewed before publication, and it still went out. I do not know how that review was done, and I draw no conclusion about Google's processes. The point is only that fluent, confident output can pass people who are paid to check it, which is precisely the failure this paper is about. **PRACTICE**

At the same time the first controlled evidence on productivity has appeared. A study released on 13 February 2023 randomly assigned 95 professional developers to work with or without GitHub Copilot on a programming task, and reports that the treated group finished 55.8% faster.⁸ **PREPRINT** Results like this will, and should, drive adoption. They also set the terms of the conversation around speed, and the rest of this paper argues that speed is only half of what a board needs to know.

DEFINITION

Appropriate reliance means accepting AI advice when it is right and rejecting it when it is wrong. It is a property of behaviour over many decisions, and it can be measured.

CHAPTER TWO

Trust is the wrong target

Much of the public discussion of AI asks whether people trust it. The more useful question is whether they rely on it in the right cases, and that is a question about behaviour that can be observed.

The vocabulary matters here, because loose terms produce loose governance. In their review of trust in automation, Lee and See treat trust as an attitude, a belief about whether a system will help a person achieve their goals in a situation of uncertainty, and they treat reliance as the behaviour that trust tends to guide.¹ PEER-REVIEWED The two can come apart. A person may distrust a tool and still lean on it because they are busy, or trust it and still check its work because their role demands it. Their review argues that the design goal should be trust that matches what the automation can actually do, so that reliance is appropriate. PEER-REVIEWED

2.1 Three things that are often confused

It helps to separate three quantities that everyday discussion runs together. Trust is what people believe about the system. Reliance is what they do with its output, whether they accept it, amend it or reject it. Decision quality is whether the final decision was right. An organisation can raise trust without raising decision quality, and it can raise reliance while decision quality falls, if people come to accept wrong advice they would once have caught.

Lee and See also argue that trust needs to be calibrated at a fine grain.¹ PEER-REVIEWED A tool can be excellent at one kind of task and poor at a neighbouring one, and a user whose trust is uniform across both will over-rely in one place and under-use the tool in the other. Generative systems make this harder, since the same interface handles tasks of widely differing difficulty with the same confident tone. My own inference is that fine-grained calibration of this kind depends on the user knowing the domain well enough to sense where the tool's competence ends. ECAVEO

2.2 What appropriate reliance looks like

The four possible outcomes of a single assisted decision can be set out simply, and the table below does so. Two outcomes are good and two are errors, and the errors differ in kind.

AI advice	Person accepts	Person rejects
Correct	Appropriate reliance; the benefit of the tool	Under-reliance; value lost, usually visible
Incorrect	Over-reliance; error passes, often invisible	Appropriate rejection; the expert as a safeguard

TABLE 1 The four outcomes of an AI-assisted decision. Source: author's framing after Lee and See (2004).

Most adoption programmes attend to the top row. Under-reliance shows up in usage figures, so managers push against it. The bottom row is where the competence paradox lives, and it is much harder to observe, since an over-relied error looks exactly like a correct decision until its consequences arrive. The bottom-right cell is the one that justifies keeping a skilled person in the process at all, and it is the cell that routine reliance erodes. **ECAVEO**

Appropriate reliance can be measured only if the organisation knows, for a sample of cases, whether the advice was right and what the person did with it, and the ordinary flow of work does not produce that information.

CHAPTER THREE

What automation research already shows

None of this is new in principle. Human factors researchers have studied the costs of capable automation for about 40 years, and their findings transfer uncomfortably well.

Generative AI is new in what it can produce, since no earlier tool could draft a legal memorandum on request. The human side of the problem is old and well studied, and ignoring that research means relearning its lessons expensively.

3.1 The ironies of automation

Lisanne Bainbridge's short paper on the ironies of automation, published in 1983, remains the best starting point.¹³ **PEER-REVIEWED** Its argument, compressed, runs as follows. Designers automate the tasks they can, which leaves the human operator with the tasks that could not be automated and with the job of monitoring the automation and taking over when it fails. Monitoring a system that rarely fails is something people do badly. Taking over in an abnormal situation demands more skill than routine operation, yet an operator who has spent months only monitoring has had little recent practice at it. The more reliable the automation, the less prepared the human is for the moment they are needed. **PEER-REVIEWED**

Bainbridge was writing about process control and flight decks, yet the structure of her argument transfers directly to the office. A knowledge worker whose drafts are written by a tool becomes a monitor of drafts, and a monitor of drafts who rarely finds a problem is in the position her paper describes. **ECAVEO**

3.2 Complacency and automation bias

The later literature gave names to two related failures. Automation complacency is a lapse in monitoring, where a person fails to notice a problem because they assume the system is working. Automation bias is a decision error, where a person follows an automated recommendation that is wrong (a commission error) or fails to act because the system gave no prompt (an omission error). Parasuraman and Manzey reviewed the empirical studies of both and drew several conclusions that bear directly on professional work.² **PEER-REVIEWED**

Their review found complacency most likely under multiple-task load, when other work competes for attention, which describes most professional roles on most days. It found both complacency and automation bias in expert as well as naive participants, and reported that automation bias could not be prevented by training or instructions, and that it could affect teams as well as people working alone. They concluded that the two phenomena overlap and that attention plays a central part in both.²

PEER-REVIEWED

ORIGIN

Bainbridge's paper runs to five pages. Its core point, that automation can make the human's remaining job harder, is the thread that runs through this whole chapter.

The finding about experts matters because organisations often assume the opposite. The comfortable belief is that senior people will catch errors because they know the domain. The review suggests that expertise does not by itself protect against automation bias, even before any erosion of that expertise over time. An organisation that relies on seniority as its safeguard is relying on something the evidence does not support. [PEER-REVIEWED](#)

A more recent literature review from Microsoft Research, covering about 60 papers, applies the same concerns to AI systems specifically.¹⁴ [PREPRINT](#) It treats over-reliance as the thing that stops users from benefiting from what AI does well while also over-seeing what it does badly, and it surveys how over-reliance arises, how it can be measured and how its effects might be reduced.

3.3 Why explanations are not a cure

The most common proposed remedy is explanation, on the theory that people who see the system's reasoning can judge whether to accept it. The evidence to date is discouraging.

Bansal and colleagues ran mixed-method user studies on three datasets in which an AI of roughly human accuracy helped participants, sometimes with explanations and sometimes without.⁴ [PEER-REVIEWED](#) They observed that the human and AI together did better than either alone, which is the result everyone hopes for. Explanations did not add to that gain. They increased the chance that people accepted the AI's recommendation, whether or not it was correct. In other words the explanation worked as persuasion, and it persuaded people towards wrong answers as readily as towards right ones.

Buçinca, Malaya and Gajos approached the problem from dual-process theory, reasoning that people rarely engage analytically with each recommendation and instead form general habits about when to follow the AI.¹⁵ [PEER-REVIEWED](#) In an experiment with 199 participants they tested three cognitive forcing designs, which compel people to think before or while seeing the AI's suggestion, against two simple explanation designs and a baseline with no AI. Cognitive forcing reduced over-reliance compared with the simple explanation designs. There were two costs. Participants gave the least favourable ratings to the designs that reduced over-reliance most, and the benefit was greater for people with a higher need for cognition, meaning those who already enjoy effortful thinking.¹⁵ [PEER-REVIEWED](#)

That pair of costs matters for any organisation, since the interventions that work are the ones users like least, and they work best for the people who least need them. A tool procured on user satisfaction scores will tend to select against the features that protect decision quality. [ECAVEO](#)

Vasconcelos and colleagues offer a partial reconciliation.⁵ [PREPRINT](#) They argue that people choose strategically whether to engage with an explanation, weighing the effort of checking against the effort and reward of simply relying on the AI. Across five studies with 731 participants, using a maze task, they found that task difficulty, explanation difficulty and monetary reward all affected over-reliance, and that explana-

tions could reduce over-reliance when they made verification cheaper. Their framing suggests that some earlier null results may reflect explanations that did not lower the cost of checking enough to matter. **PREPRINT**

Clinical evidence points the same way. Gaube and colleagues gave physicians chest X-rays with diagnostic advice, some of it inaccurate, and labelled some advice as coming from an AI system when all of it had in fact come from human experts.³ **PEER-REVIEWED** Radiologists rated advice lower when it appeared to come from AI, while physicians with less task expertise did not. Diagnostic accuracy was significantly worse when participants received inaccurate advice, whatever its purported source. Scepticism towards AI as a source, then, did not protect even the specialists from bad advice. **PEER-REVIEWED**

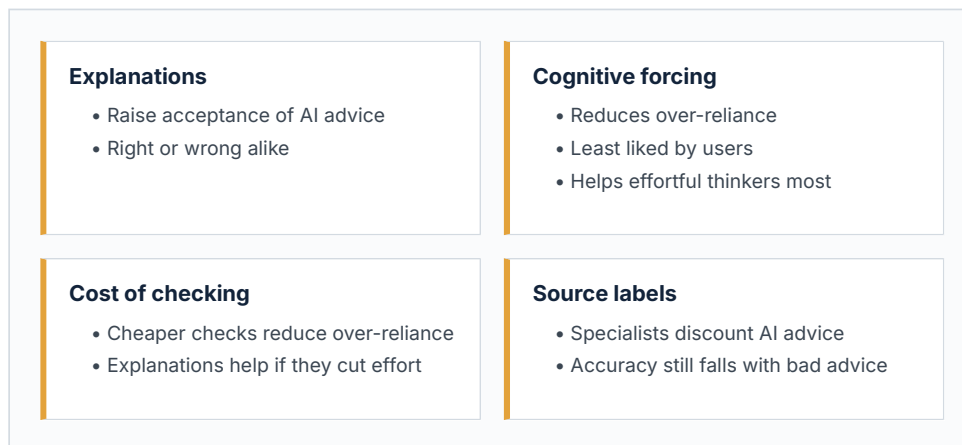
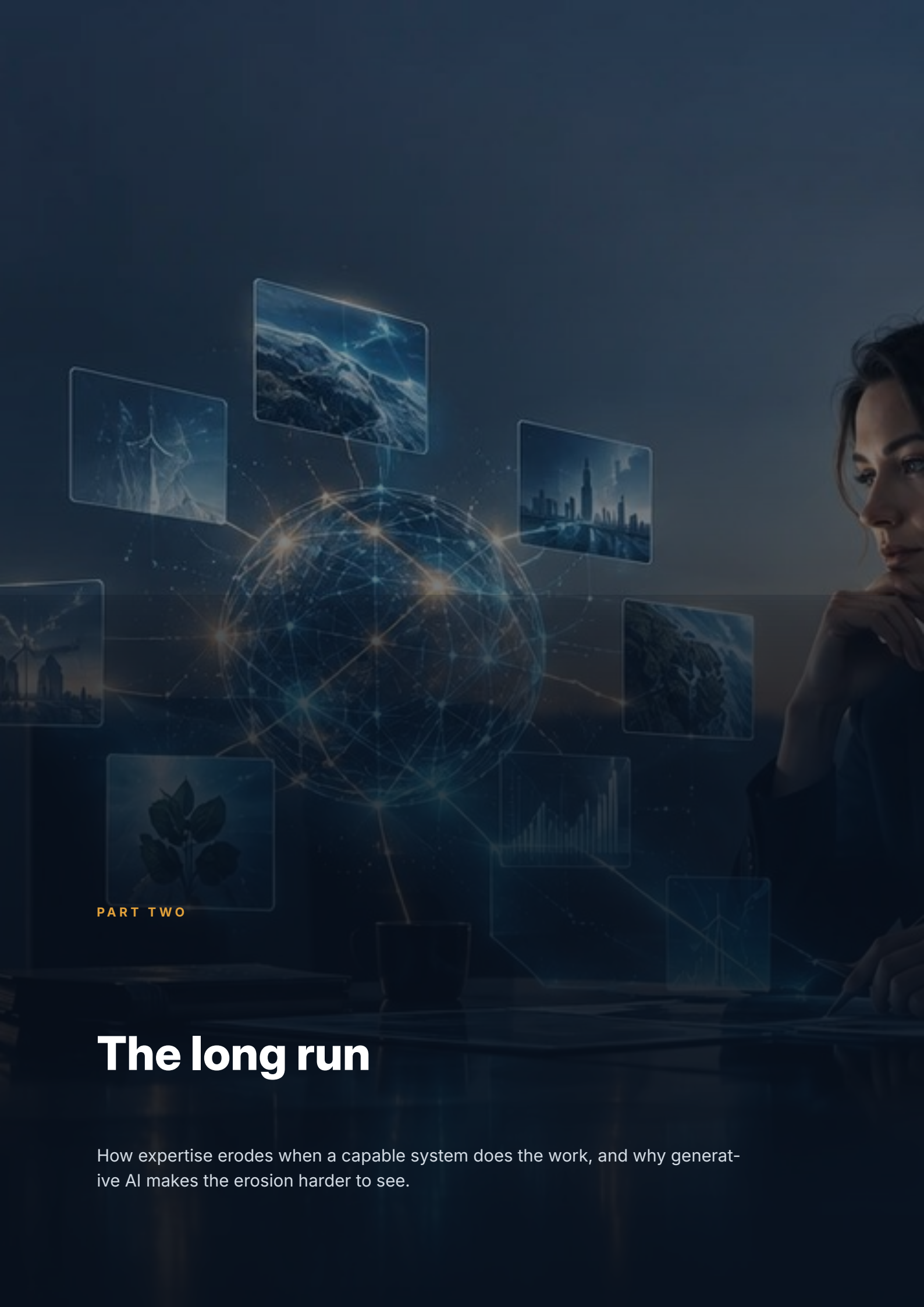


FIGURE 1 What the decision-support studies suggest Source: Bansal et al. (2021); Buçinca, Malaya and Gajos (2021); Vasconcelos et al. (2022); Gaube et al. (2021).

Taken together these studies suggest that the design of the interaction can reduce over-reliance, and that explanation alone does not. They share a limitation, though, which is the subject of the next part. Almost all of them measure a single session.



PART TWO

The long run

How expertise erodes when a capable system does the work, and why generative AI makes the erosion harder to see.

CHAPTER FOUR

The long-run problem of skill erosion

The studies of over-reliance mostly observe people for an hour. The competence paradox unfolds over months and years, and the evidence for that timescale comes from aviation, medicine and a small number of workplace studies.

A single-session experiment can show that people accept wrong advice, but it cannot show what happens to a professional's judgement after two years of accepting mostly right advice. The research on that second question is thinner, though what exists is consistent.

4.1 Evidence from the cockpit

Aviation offers the closest thing to a long-run natural experiment, because airline pilots have flown highly automated aircraft for decades. Casner and colleagues asked 16 airline pilots to fly routine and non-routine scenarios in a Boeing 747-400 simulator, varying the level of automation and probing what the pilots were thinking about as they flew.⁶ PEER-REVIEWED

The result is more precise than the popular fear. Pilots' instrument scanning and manual control skills were mostly intact, even where the pilots reported practising them rarely. The difficulties lay in the cognitive work of manual flight, such as tracking the aircraft's position without a map display, deciding which step of the route came next and recognising instrument failures. There the researchers observed more frequent and significant problems. Performance on these cognitive tasks was associated with how often pilots engaged in thought unrelated to the task while the automation flew. The authors concluded that retention of those cognitive skills may depend on how actively pilots stay engaged in supervising the automation.⁶

PEER-REVIEWED

Two features of this study carry over to knowledge work. The first is that the skills that faded were cognitive, the kind of situational understanding that lets a person know where they are and what should happen next. Most professional expertise is of that kind. The second is the link to engagement. The pilots who let their minds wander while the system worked were the ones who struggled when asked to do the thinking themselves, which suggests that the manner of supervision, alongside its duration, shapes what a person retains. (A small study, of course, and simulator performance is not line performance.) PEER-REVIEWED

4.2 Clinical and knowledge work

Medicine has raised the same concern with less experimental evidence. Hoff interviewed 78 primary care physicians in the United States about two work innovations, electronic medical records and clinical guidelines.¹⁶ PEER-REVIEWED The physicians de-

scribed deskilling outcomes that included decreased clinical knowledge, decreased confidence in making clinical decisions, decreased patient trust and more stereotyping of patients. Hoff also found that physicians took an active part in the deskilling through the adaptive choices they made to fit the innovations into their daily work, and that the business model of primary care made those effects worse.¹⁶

PEER-REVIEWED

That second finding matters most for a board. In Hoff's account, deskilling follows from the professionals' own sensible local choices, made under time pressure and rewarded by the surrounding system, which is the pattern in the invented scenario above. **ECAVEO**

Writing in JAMA, Cabitza, Rasoini and Gensini warned that machine learning decision support in medicine could have unintended consequences, deskilling among them, alongside the risk that such systems neglect clinical context and the uncertainty inherent in medical data.¹⁷ **PEER-REVIEWED** Their piece argues from principle and clinical experience, and it is best read as a warning from informed practitioners.

The most direct evidence from ordinary knowledge work comes from a case study by Rinta-Kahila and colleagues.⁷ **PEER-REVIEWED** They studied an IT service firm whose managers decided to discontinue a software system that had automated fixed assets accounting and reporting, apparently effectively. The authors describe how the automation had produced latent deskilling among the accountants, which became apparent only when the system was withdrawn and caused disruption to daily work and organisational processes. They also document how the employees and the firm recovered. Their central suggestion is that the effect of automation on skills may not be obvious while the system operates, and that discontinuation brings it into view.⁷

PEER-REVIEWED

Deskilling is invisible while the system works. Lost skill, the accounting case suggests, rarely announces itself. It surfaces when the tool is withdrawn, fails or meets a case it cannot handle. An organisation that waits for that moment to learn what its people can still do has chosen the most expensive way to find out.

4.3 The expertise that catches errors

Pulling these strands together, I would describe the mechanism as a loop with five stages, set out in the figure below. It is my synthesis of the evidence, which none of the cited authors proposed in this form, and each link in it rests on evidence of different strength. **ECAVEO**

1 Good output

The tool is right on most ordinary tasks, so checking rarely finds anything.

2 Checking fades

Busy people reduce checking where it seldom pays, as attention research predicts.

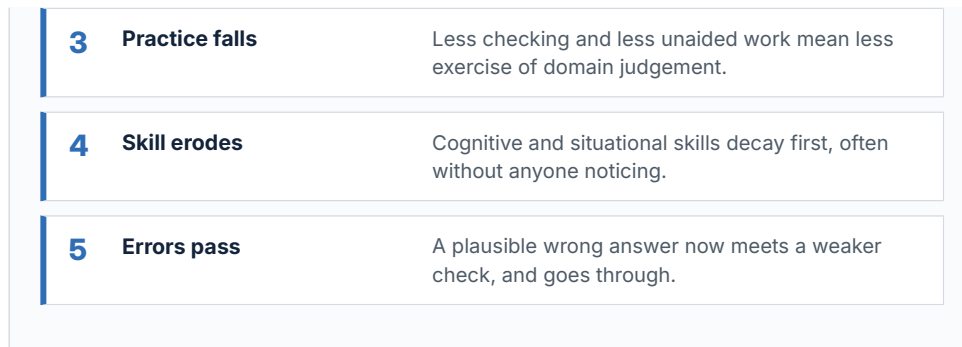


FIGURE 2 The erosion loop by which routine reliance can weaken the expertise that oversight depends on. Source: author's synthesis of Bainbridge (1983), Parasuraman and Manzey (2010), Casner et al. (2014) and Rinta-Kahila et al. (2018).

The loop has an awkward property, which is that every stage except the last looks like success. Output is good, people are efficient, and work gets done faster. Only the final stage produces a visible event, and by then the capacity to have caught it has been lost for some time. I would add one further point, which goes beyond anything the cited studies test. The expertise that catches errors is largely tacit, built up through years of doing the work and noticing when something feels wrong, and tacit knowledge is the kind least likely to be written down and most likely to fade when the work that exercised it stops. **ECAVEO**

A related question I have not resolved is whether the loop runs differently for people who never built the expertise in the first place. A junior professional who learns the job with a capable tool from the first day may never develop the judgement that a senior colleague is at risk of losing. I would call that an unfilled gap, and I suspect it will prove the harder problem for professions that train through apprenticeship. I know of no evidence on it yet. **EVIDENCE GAP**

CHAPTER FIVE

Generative AI raises the stakes

Earlier decision aids were narrow and their errors had a recognisable shape. Generative systems are fluent across open-ended tasks, and their errors read like competent work.

Most of the research reviewed so far concerns systems with a bounded job, such as classifying an image, flagging a transaction or proposing a diagnosis from a fixed list. A person working with such a system knows what kind of output to expect and, in time, what kind of error. Generative AI changes both.

5.1 Fluent output across open-ended work

Richard and Daniel Susskind argued in 2015 that increasingly capable systems would change how professional expertise is produced and delivered, and that much professional work would be broken into tasks, some of them handed to machines.¹⁸

SCHOLARLY The tools released since November 2022 have made that argument concrete for much writing, analysis and coding.

Three properties of these tools matter for the competence paradox. The first is breadth. A single interface drafts, summarises, calculates, codes and advises, so the calibration problem Lee and See describe now spans many tasks at once, each with its own hidden rate of error.¹ **PEER-REVIEWED** The second is fluency. OpenAI's own announcement of ChatGPT states that it sometimes produces answers that sound plausible but are incorrect or nonsensical.¹⁰ **INDUSTRY** A narrow classifier that is wrong produces a wrong label, which a trained eye may spot, whereas a generative system that is wrong produces a well-organised paragraph, and the surface cues that normally signal carelessness are absent. The third is that the output is the work product itself. The analyst in the scenario did not receive a recommendation and then write the commentary. She received the commentary, so the only point at which her expertise could act was the review, and review is the stage that Bainbridge and the complacency literature show people perform least well.¹³ **PEER-REVIEWED**

My inference from these three properties, which goes beyond what any study has yet tested for generative tools, is that the conditions for automation bias are stronger here than in most of the settings where it has been measured. **ECAVEO** The Bard episode proves nothing on its own, but it fits the pattern. An error about the history of astronomy, a subject where a moment's checking would have helped, passed into published promotional material.¹² **INDUSTRY**

5.2 Why productivity studies miss the risk

The Copilot study is careful work.⁸ **PREPRINT** Of the 95 developers recruited, 45 were assigned to the treated group and 50 to the control group, and among those who completed the task the average completion time was about 71 minutes with the tool

OBSERVATION

Conversational tools also invite a social response. Lee and See note that people respond to technology socially,¹ **PEER-REVIEWED** and a system

against about 161 minutes without. The authors report a 95% confidence interval for the improvement of 21% to 89%, and find that less experienced developers, among others, appear to benefit more.⁸ **PREPRINT**

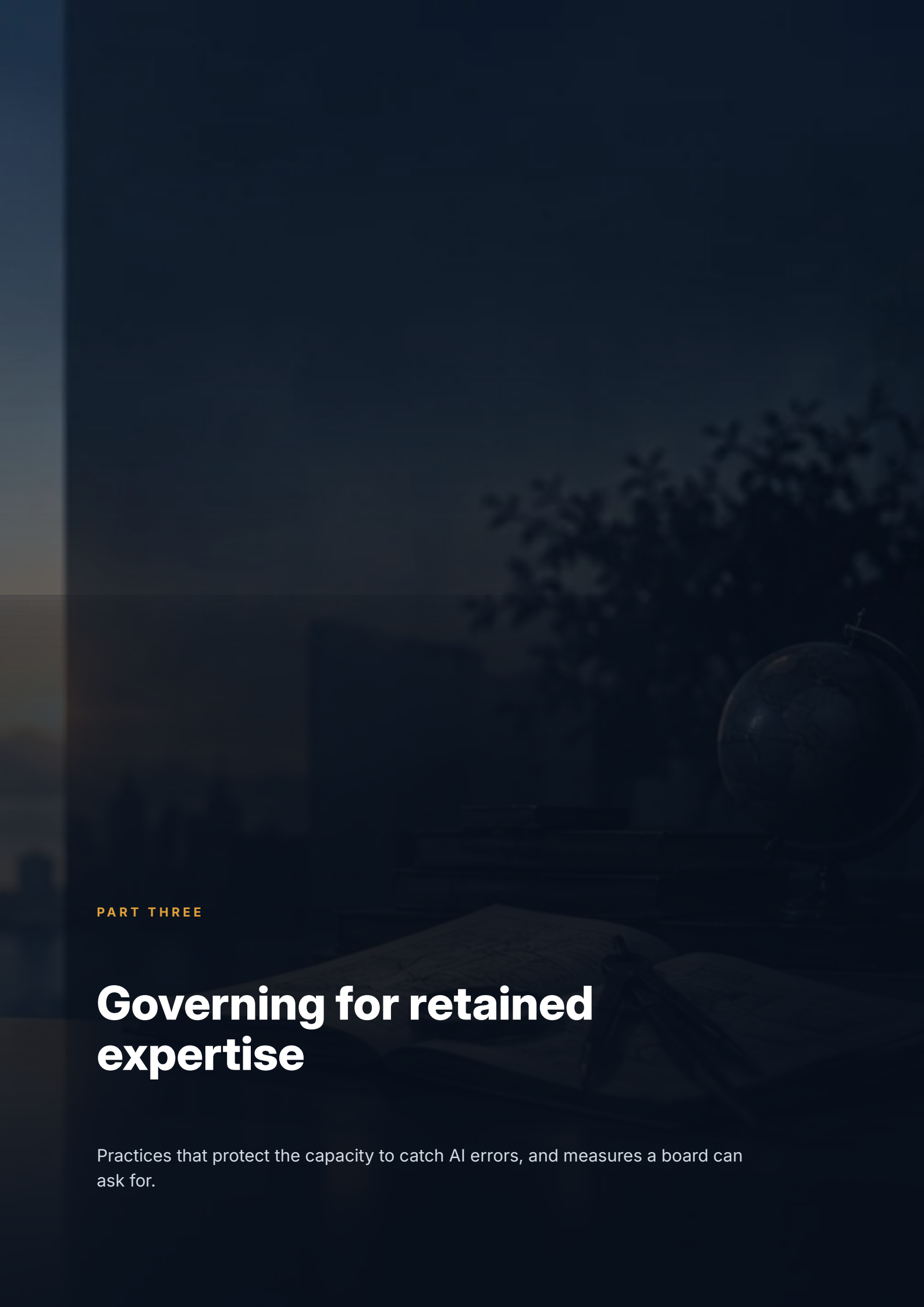
that writes in confident prose may attract trust its accuracy has not earned.



FIGURE 3 Average task completion time in the GitHub Copilot experiment, among developers who completed the task. Source: Peng et al. (2023), arXiv:2302.06590.

The design, however, answers a question about speed on a single, well-specified task with a test suite that tells the developer when the work is correct. In that setting the test suite does the checking, so the study cannot show how well developers would catch a subtle flaw that the tests did not cover. Nor can a single session show what happens to a developer's own design skill after a year of accepting suggestions. The authors claim nothing more. **EVIDENCE GAP**

A 55.8% speed gain is the kind of figure that drives a business case, and a business case built on it will tend to measure what the study measured.⁸ **PREPRINT** If the finding that less experienced people gain most holds up, it also creates an incentive to staff more work with less experienced people using the tool, which may be sensible for throughput and is precisely the configuration in which fewer people can catch a fluent error. That is my reading, and nobody has yet tested it. **ECAVEO**

A dark, atmospheric background image showing a desk with a globe, books, and papers. The scene is dimly lit, with a warm glow from the left side, possibly a window or a lamp. The globe is on the right, and several books are stacked in the center. Papers are scattered on the desk in the foreground.

PART THREE

Governing for retained expertise

Practices that protect the capacity to catch AI errors, and measures a board can ask for.

CHAPTER SIX

Designing for retained expertise

The research suggests that over-reliance and skill loss respond to design. An organisation can build work so that expertise is exercised on purpose.

Nothing in the evidence argues for refusing the tools, and studies show that human and machine together can outperform either alone.⁴ **PEER-REVIEWED** The task is to keep the gains. What follows is my proposal, drawn from the studies above, and it should be read as a framework to try and refine. **ECAVEO**

6.1 Four practices

I propose four practices, each aimed at a different stage of the erosion loop. The figure summarises them and the paragraphs that follow give the reasoning.

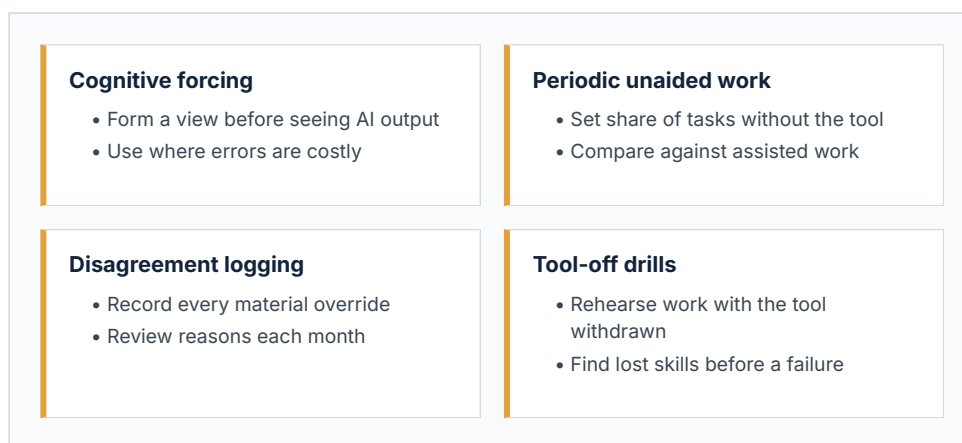


FIGURE 4 Four practices for retaining expertise in AI-assisted work. Source: author's framework, drawing on Bućinca, Malaya and Gajos (2021), Casner et al. (2014) and Rinta-Kahila et al. (2018).

Cognitive forcing comes first because it has the best direct evidence.¹⁵ **PEER-REVIEWED** In high-stakes decisions the professional should record a provisional view before seeing the AI output, or should be required to engage with the output in a structured way before accepting it. The research suggests that people will dislike this, so it should be confined to decisions where an undetected error would be costly, and it should be presented as part of the professional standard. **ECAVEO**

Periodic unaided work addresses the practice gap that Bainbridge and Casner describe. A defined share of real tasks, chosen at random, is done without the tool, and the output is compared with assisted work on similar tasks. The comparison serves two purposes, since it keeps the skill in use and it gives the organisation a sample of what its people can do on their own. I do not yet know what the right share is, and it will differ by role. My starting suggestion is a small, fixed proportion reviewed each quarter, adjusted once the data show whether unaided quality is holding. **ECAVEO**

Disagreement logging creates the evidence the organisation otherwise lacks. Each time a professional materially overrides or corrects AI output, they record what they changed and why. The log shows where the tool fails, which improves calibration for everyone, and a falling override rate becomes a prompt for inquiry. It may mean the tool has improved, or it may mean people have stopped looking. (The log needs to be light, a line or two, or nobody will keep it.) **ECAVEO**

Tool-off drills borrow from aviation and from the accounting case. Periodically a team works through a realistic task with the tool withdrawn, on purpose and without penalty, so that any lost skill surfaces in a rehearsal, where finding it costs little.⁷

PEER-REVIEWED The drill also tests the organisation's own processes, since a team may find that nobody remembers how a manual step was once done. **ECAVEO**

6.2 What the frameworks already say

Two risk management publications appeared in the weeks before this paper closed, and both touch on human oversight. The US National Institute of Standards and Technology released version 1.0 of its AI Risk Management Framework on 26 January 2023.⁹ **STANDARD** Its GOVERN function calls for policies that define and differentiate roles and responsibilities for human-AI configurations and oversight, and its MAP function includes processes for operator and practitioner proficiency and for human oversight. An appendix on human-AI interaction notes that the degree to which people are empowered and incentivised to challenge AI output needs further study, and suggests that data on how often and why people overrule AI output in deployed systems may be useful to collect and analyse.⁹ **STANDARD**

ISO/IEC 23894, published in February 2023, gives guidance on how organisations that develop, deploy or use AI can manage the risks specific to it and integrate that risk management into their activities.¹⁹ **STANDARD**

My own reading, which goes beyond what either document says, is that both frameworks treat human oversight as a control to be defined and assigned, and neither says much about how the skill of the overseer is kept current over time. The NIST suggestion about override data is the closest either comes, and it is a good one. A board that adopts the four practices above would, on my reading, be putting flesh on the proficiency and oversight elements that the NIST framework already names. **ECAVEO**

CHAPTER SEVEN

Measures a board should ask for

What is not measured will not be protected. A board needs a few indicators of retained expertise, reported beside the productivity figures it already receives.

Vendors and managers supply adoption and output data readily, whereas the measures below are harder to collect and less flattering, so a board has to ask for them. They are my proposal, and the thresholds that should trigger concern will need to be set by each organisation once it has a baseline. **ECAVEO**

7.1 The measures

For each role where an undetected AI error would be costly, a board should ask for four measures, reported quarterly.

- The override rate, meaning the share of AI outputs that the professional materially changed or rejected, with its trend over time and the main reasons recorded.
- The error-catch rate, measured by seeding a small number of known errors into AI output under review, or by independent re-review of a sample, to estimate how often real errors are caught.
- Unaided performance, taken from the periodic unaided work and scored against the same standard as assisted work.
- The results of the most recent tool-off drill, including any process or skill gaps it revealed and what was done about them.

The value of these measures comes from reading them beside the productivity and quality figures the board already sees. Rising output with a stable error-catch rate and steady unaided performance is the healthy pattern. Rising output with a falling override rate, a falling catch rate and weakening unaided work is the competence paradox made visible, and it calls for action before any failure occurs. **ECAVEO**

Seeded errors need care. They should be designed with the people being tested, disclosed as a practice in advance, and used only to improve the system, since ranking people on the results would distort the behaviour the measure is trying to observe. Error-catch testing is familiar in some safety-critical fields, and I expect it can be adapted to knowledge work, though I know of no published evaluation of doing so with generative tools. **EVIDENCE GAP**

7.2 Conclusion and open questions

The argument of this paper is that the value of AI assistance depends on a human capability that the assistance itself can weaken. Appropriate reliance, meaning the acceptance of good advice and the rejection of bad, is the right target. It depends on expertise, and the evidence from human factors, aviation, medicine and one careful workplace study suggests that expertise fades when capable automation removes the

practice that sustains it, and that the fading is hard to see until the automation fails.

^{2,6,7} **PEER-REVIEWED** Generative AI, fluent across open-ended work, seems likely to strengthen the conditions for that pattern, and the productivity studies now appearing are not designed to detect it. **ECAVEO**

The practical answer is to keep adopting while designing work so that expertise is exercised on purpose, and to measure retained expertise as seriously as output, starting with the practices and measures proposed here and changing them as organisations learn.

Several questions remain open, and I would welcome evidence on any of them.

- What share of unaided work is enough to maintain expertise in roles that rely heavily on AI, and does it differ between drafting, analysis and decision roles?
- Should professional bodies require evidence of unaided competence as part of revalidation or continuing professional development?
- How can an organisation detect that expertise has eroded before a failure exposes it, in fields where errors surface slowly?
- What happens to people who learn a profession with a capable tool from the start, and never build the judgement that others are at risk of losing?

The last of these may matter most over a decade, and it is the one on which, as far as I can find, there is at present no evidence at all.

END MATTER

References

1. Lee, J.D. and See, K.A. (2004) Trust in automation: designing for appropriate reliance. *Human Factors* 46(1), 50-80.
2. Parasuraman, R. and Manzey, D.H. (2010) Complacency and bias in human use of automation: an attentional integration. *Human Factors* 52(3), 381-410.
3. Gaube, S. et al. (2021) Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digital Medicine* 4, 31.
4. Bansal, G. et al. (2021) Does the whole exceed its parts? The effect of AI explanations on complementary team performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*, 1-16.
5. Vasconcelos, H. et al. (2022) Explanations can reduce overreliance on AI systems during decision-making. arXiv: 2212.06823.
6. Casner, S.M., Geven, R.W., Recker, M.P. and Schooler, J.W. (2014) The retention of manual flying skills in the automated cockpit. *Human Factors* 56(8), 1506-1516.
7. Rinta-Kahila, T., Penttinen, E., Salovaara, A. and Soliman, W. (2018) Consequences of discontinuing knowledge work automation: surfacing of deskilling effects and methods of recovery. *Proceedings of the 51st Hawaii International Conference on System Sciences*, 5244-5253.
8. Peng, S., Kalliamvakou, E., Cihon, P. and Demirer, M. (2023) The impact of AI on developer productivity: evidence from GitHub Copilot. arXiv:2302.06590, 13 February 2023.
9. National Institute of Standards and Technology (2023) *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. Gaithersburg, MD: NIST, 26 January 2023.
10. OpenAI (2022) ChatGPT: optimizing language models for dialogue. OpenAI blog, 30 November 2022. Available at: openai.com/blog/chatgpt.
11. Mehdi, Y. (2023) Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web. Official Microsoft Blog, 7 February 2023.
12. Reuters (2023) Alphabet shares dive after Google AI chatbot Bard flubs answer in ad. Reuters, 8 February 2023.
13. Bainbridge, L. (1983) Ironies of automation. *Automatica* 19(6), 775-779.
14. Passi, S. and Vorvoreanu, M. (2022) *Overreliance on AI: Literature Review*. Microsoft Research, MSR-TR-2022-12.
15. Buçinca, Z., Malaya, M.B. and Gajos, K.Z. (2021) To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5(CSCW1), Article 188.
16. Hoff, T. (2011) Deskilling and adaptation among primary care physicians using two work innovations. *Health Care Management Review* 36(4), 338-348.
17. Cabitza, F., Rasoini, R. and Gensini, G.F. (2017) Unintended consequences of machine learning in medicine. *JAMA* 318(6), 517-518.
18. Susskind, R. and Susskind, D. (2015) *The Future of the Professions: How Technology Will Transform the Work of Human Experts*. Oxford: Oxford University Press.
19. International Organization for Standardization and International Electrotechnical Commission (2023) *ISO/IEC 23894:2023 Information technology. Artificial intelligence. Guidance on risk management*. Geneva: ISO, February 2023.

END MATTER

About the author

Paul Forrest is the founder of Ecaveo. He advises boards and leadership teams on the adoption and governance of artificial intelligence, drawing on a background in business process reengineering and on hands-on work building machine learning systems, including small, domain-specific language models. He has served as a special adviser to UK government on the Global Entrepreneur Programme and works as a board member and business adviser.

He is also a qualified athletics coach who coaches sprint, hurdle and endurance athletes as a volunteer, and much of his thinking about how people learn, decide and improve comes from the track.

This paper is part of the Ecaveo Working Papers, AI Adoption: The Higher-Order Issues. The views expressed are the author's own.

ecaveo.com

The Competence Paradox

Ecaveo Working Papers, AI Adoption: The Higher-Order Issues
Paper two of six · 21 February 2023