SAFWAN ALMODHAYAN



OPINION AND ANALYSIS

Editorial date: October 29, 2026

The Apprentice Whose Mistakes Disappeared

> The problem is not that a machine has mastered what a beginner used to do. It is that some tasks that look easy to eliminate were, invisibly to payroll, the school from which experts graduated.

On January 4, 2013, the US Federal Aviation Administration issued a one-page safety alert titled “Manual Flight Operations.” It did not celebrate a romantic return to the controls or call autopilot a bad invention. It acknowledged that automated flight systems had improved safety, workload management, and operational precision. Then it identified a quieter problem: continuous use of those systems does not necessarily reinforce the knowledge and skills needed for manual control, and may weaken a pilot’s ability to recover quickly when the aircraft leaves the desired state.1

Its recommendation was precise: suitable opportunities for manual practice in low-workload conditions, while continuing to use automation when safety and precision require it. In 2017 the agency again emphasized developing and maintaining manual proficiency through training and operations, without turning the issue into a contest measured by minutes spent defying automation.2

The concern did not rest on one page. A 2013 report by the Flight Deck Automation Working Group addressed pilot training and flight-path management in a broader technical review, confirming that operating and training design cannot be separated from the characteristics of the automated system.13

What merits attention is not fear of a machine flying an aircraft. It is the question the alert placed where most management reports would rather not put it: **if the machine performs ordinary work well, where will humans acquire the expertise needed for the extraordinary moment?**

In a software office, contract-review team, or diagnostic laboratory, the question seems less dramatic. There is no cockpit or altitude. Yet something similar may disappear: hours of safe error, correction, another attempt, and explanation of why the first answer failed. Company accounting sees slow cost in those hours. The economy of a profession sees a compounding asset.

When mistakes disappear from an apprentice's work, some of what would have made that apprentice an expert may disappear with them.

A Small Task With a Large Function

It is easy to explain why a company wants to automate junior tasks. A new engineer writes routine tests; a legal analyst searches precedents; a bank employee reviews similar documents; a trainee doctor prepares an initial case summary under specialist supervision. If all of this can be done faster and more cheaply, who would defend slowness for its own sake?

But an institutional task is not always what its job description suggests. It may have two functions.

The first is producing a useful output today: working software, a clear memorandum, or an accurate review. The second is producing capacity in the person who did it: noticing a pattern, forming judgment, and learning when a plausible answer is wrong.

The first is sold to the customer immediately; the second shows its value years later. A manager sees the first on a weekly dashboard. The second may become visible only when an experienced employee leaves and the team discovers that nobody learned how to think as they did.

We must distinguish three things that the language of transformation often mixes together: **task completion, skill acquisition, and skill retention**. A system may succeed at the first without adding anything to the second. A tool may help someone acquire knowledge now without telling us whether it remains six months later. A worker may master a task for years and then lose proficiency in one part through lack of practice. Each phenomenon has a different mechanism and body of evidence.

In the economics of human capital, Yoram Ben-Porath discussed as early as 1967 how people allocate time and resources to producing future capacity, not only current income.3 In professional life this investment often takes an invisible form: some time that appears "unproductive" by an immediate commercial measure creates tomorrow's productivity.

If a company removes all that time, it has not merely shortened a process. It may have changed the production function of expertise.

Machine-assisted task completion is not imaginary. In a controlled experiment first released as a draft in 2023, developers using GitHub Copilot finished a web-server programming task 55.8 percent faster than a group without it. But the experiment measured time on a particular task; it was not designed to measure what participants learned after closing the tool.14 The distinction between those questions deserves a fair hearing.

The Experiment That Separated Output From Learning

In a preliminary paper released on January 28, 2026, Judy Hanwen Shen and Alex Tamkin asked what happens to skill acquisition when developers use an AI assistant while learning a new software library. *How AI Impacts Skill Formation* is an **unreviewed research draft**, a status that matters when weighing its result.4

The researchers chose the asynchronous Python library Trio. Participants were not complete programming novices: each had more than one year of Python experience and many had several, but none had used the library under study. This detail reverses the easy reading. The experiment did not compare a child learning to code with a machine that already could; it compared relatively experienced people confronting a new skill.

The main study included 52 participants divided equally. Twenty-six received a coding assistant and 26 worked without one. All faced time-limited tasks, followed by an independent test of conceptual understanding and the ability to read and debug code.

The striking result was not merely lower test performance with assistance, but that **completion and learning moved in different directions**. The AI group averaged about 50 percent on the skills test, compared with roughly 67 percent for the other group—a statistically significant difference of 17 percentage points in this test. The main experiment did not, meanwhile, establish a significant improvement in average task-completion time.4

Because precision matters more than the force of the number, four limits belong beside it. The sample was relatively small; exposure was short; the test closely followed training; and the domain was one library, not general programming ability or the fate of programming over a decade. The AI group all completed the second task, while four control participants failed to do

so within the time limit, which means the definition of “productivity” can itself change the reading. This is not proof that AI corrupts programmers. It is experimental evidence of **a possible trade-off in learning design**.

Usage logs also distinguished participants who delegated the whole task from those who asked for a concept or an error to be explained and continued thinking. More effortful interaction patterns were associated with better learning results. But that analysis was observational within the assisted group; it cannot by itself establish a causal effect for every pattern.

The lesson is more exact than “use less AI.” Ask: **what part of the thinking remained with the learner after use?**

How Does an Error Become a Training Device?

Imagine two developers meeting the same error message. One pastes it into a language assistant and receives a working fix. The other consults the library documentation, tests two hypotheses, sees one fail, and discovers that the problem lies not in the line’s syntax but in the order of asynchronous events. Both end with correct code. Their cognitive histories differ.

This is a hypothetical illustration, not an account of a participant in the Shen and Tamkin study. It explains a mechanism rather than generalizing a result.

In the second case, educational psychology would describe retrieval practice followed by feedback comparing expectation with reality. Experiments by Henry Roediger and Jeffrey Karpicke in 2006 showed that testing retrieval can improve later retention beyond what immediate performance may suggest under some experimental conditions.[5] The claim is not that every failure helps. It is that attempting retrieval and checking the result are not educational waste that can safely be removed by default.

In work on “productive failure,” Manu Kapur described how initial attempts that do not reach the right answer can, when designed for learning and followed by consolidation and explanation, prepare deeper understanding.[6] Turning this into “let the employee make mistakes” would be dangerous. Error helps when its cost is contained, explanation and another attempt are available, and the learner can connect a failed attempt with a more accurate principle. Throwing a trainee into a dangerous case without supervision is not learning; it is bad management dressed as pedagogical courage.

Expertise is not the number of errors someone has made. It is how much their mental model changed after they understood why they were wrong.

An Old Irony in a New Machine

Decades before language assistants, Lisanne Bainbridge published “Ironies of Automation” in 1983. One central idea was that automation may push people away from daily practice and then demand their intervention in the hardest cases. The machine handles the ordinary and leaves the exceptional to the human, even though the ordinary supplied part of the training needed for the exceptional.[7]

The analogy is incomplete. Flying is a complex sensory and motor skill, contract reading is institutional interpretation, and debugging code is logical and empirical. These abilities do not decay at the same rate, and a vivid metaphor cannot establish one effect across them. Automation may also grow new abilities, such as specifying requirements and reviewing model outputs.

The hypothesis is not that all manual practice is virtuous. It is that **the kind of practice disappearing and the kind of judgment later required must be studied together**.

A team led by Brooke Macnamara revisited the possibility that automated assistance accelerates skill loss in a 2024 scientific article. Its value is in defining a research problem and possible mechanisms, not in experimentally proving general decline in every occupation.[12]

An experiment by Stephen Casner and colleagues, published in *Human Factors* in 2014, supplies a necessary correction to easy pessimism. Using a Boeing 747-400 simulator with 16 pilots, the researchers found that aspects of basic manual performance remained robust while certain cognitive tasks appeared more vulnerable to decline.[8] If the law were “all automation erases all skill,” these differences would not appear. We need to know which element eroded: movement, attention, diagnosis, systems understanding, or coordination under pressure.

If a software screen always presents the correct answer, a manager may think errors have been eliminated. Perhaps only the occasions that revealed how a beginner thinks have disappeared.

A Doctor Sees More—and Then What Without the Screen?

In a 2025 observational study in *The Lancet Gastroenterology & Hepatology*, researchers followed the introduction of AI assistance at four colonoscopy centers in Poland. They did not measure only how well the system worked when present. They observed conventional examinations conducted **without automated assistance** before and after its introduction.

The comparison included 1,443 unassisted examinations: 795 before introduction and 648 after. The adenoma detection rate fell from 28.4 to 22.4 percent, a six-percentage-point difference in the sample. The researchers stated the result cautiously: sustained exposure to

assistance may indicate a change in endoscopist behavior.9

Excitement should be resisted. This was not a randomized experiment proving that the machine caused lost skill. Patient mix, center behavior, or other factors may have changed across periods, and adjusted analysis cannot remove every competing explanation. Detection rate is not the sole equivalent of clinical competence or patient outcome. The study does offer a signal deserving independent study: what happens to human observation when a tool alerts it to fine details? An expert activity may become an awaited result.

Medicine cannot sacrifice a present patient's safety to train a future doctor. General enthusiasm for manual work does not solve this problem. Alternatives may include image sets with known outcomes, blinded simulation, or collaborative review of images flagged by the system, while monitoring independent competence without withholding tools proven beneficial to patients.

Nor does all evidence point in one direction. A systematic review and meta-analysis of AI-augmented teaching for simulated surgical skills, published in June 2026, quantitatively analyzed four studies and 268 participants. AI-augmented instruction was comparable to expert instruction on several measures, with a small advantage on one practical-skills measure. Evidence certainty was low, and the researchers found no basis for replacing human teachers.10

This counter-result is valuable. The question is not "does the tool teach or damage?" It is: what teaching method, which skill, what feedback, and who decides that a trainee is ready for practice in which repeated error is unacceptable?

One Hint Can Equal a Full Answer in Production, but Not in Learning

During the 2023–2024 school year, researchers ran a field experiment with nearly one thousand Turkish high-school students, testing two kinds of AI assistance in mathematics. Results were published in *PNAS* in 2025.11

The first was a general chat-like interface using GPT-4, called "GPT Base." The second, "GPT Tutor," used teacher guidance to provide hints and account for common errors and correct exercise solutions, instead of quickly giving the final answer. Other students used neither assistant.

During practice, the first group's performance improved by a relative average of 48 percent over the control group; the guided-assistant group improved by 127 percent. A marketing presentation satisfied with that result might stop there. The experiment did not.

When tested without assistance, users of the general assistant performed a relative 17 percent below the control group. The teaching-oriented assistant reduced that negative effect until the difference from control was no longer statistically significant, but it did not establish superior independent-test performance either.

A **relative 17 percent decline in the mathematics experiment** must not be confused with the **17-percentage-point difference on the Trio programmers' average test score**. A similar numeral does not make the measure, population, or training period the same.

The more important result is that learning does not consist merely of obtaining the right answer. A complete answer removes uncertainty at once; a hint asks the learner to reduce it. In one educational setting, a system giving less information may preserve a larger educational function.

This is not a defense of deliberate ignorance. A useful hint requires precise knowledge of a learner's position and difficulty, followed by trustworthy correction. The tutor version was not merely a gentler prompt; it used problem-specific preparation and real teacher input. "Do not give the answer" is easy to copy into a strategy document. Building a system that knows when to stop hinting before the learner circles an unexplained error is harder.

The Economy Missing From the Cost Sheet

Suppose an institution has 100 monthly review tasks, each requiring two junior hours before automation and half an hour after assistance. This is **a hypothetical calculation**, not a company measurement. The accountant will see 150 hours "saved," and they may indeed deserve saving.

But what if some old hours produced something else: the ability to identify an exceptional case, build an argument from conflicting sources, or recognize a defective document? Unrecorded training hours may become a future cost through dependence on a few senior employees, audit errors, or the need to buy experts from a market whose firms have stopped producing them.

An illustrative relationship—not an empirically estimated equation—can express the idea:

$$H_{t+1} = (1 - \delta)H_t + \alpha\, P_t\, F_t.$$

Here H_t is the skill stock of a person or team; δ is its rate of loss when unused; P_t is the amount of **relevant practice**; F_t is feedback quality and opportunity to understand error; and α is the extent to which practice converts into learning. Not every work hour is training, and not every correction helps. An assistant may raise F_t and deepen practice, or lower P_t by doing everything instead. It may change the required skills altogether.

This framing distinguishes three cases:

- Automating repetitive transcription that produced no real learning: performance can rise without material loss in H.
- Automating a task in which the trainee found causes and tested hypotheses: risk appears if practice shrinks without replacement.
- An assistant that explains, generates tests, and exposes errors without monopolizing the solution: both P_t and F_t may rise even as total time falls.

The point is not to pay employees to reinvent the wheel. In some professions, reinventing a small piece of it in a simulator under supervision is what teaches someone why it might one day fail.

Call this gap an **expertise reproduction deficit**. It is not a published theory or official standard, but a diagnostic concept proposed here: expected demand for independent expert judgment grows faster than the organization's ability to produce people capable of exercising it without assistance.

Operationally, it can be measured as a gap between the departure or transfer rate of current experts and the number of trainees reaching an independent competence threshold in a specified period. The testable prediction is that if diagnostic practice is removed without an effective educational substitute, new cohorts will show a widening gap on tests of independent judgment even while their AI-assisted productivity improves.

The concept fails if work changes until the old judgment is no longer needed, or if the tool creates a better learning route. Demand for the ability itself must be measured, not a skill worshipped after losing its function.

The Board Is Asking the Wrong Question

Professional automation projects are usually presented with licence cost, employee hours, tasks completed, and detected errors. These are legitimate figures. The table needs another column less suited to a conference screen: **what skills were acquired while performing this task, and who will acquire them after it is automated?**

Initial contract review might teach recognition of a clause carrying hidden risk; an ordinary bank case might train the connection between customer identity and transaction purpose; searching an industrial record might teach the distinction between a plausible cause and mere concurrence. These are possible mechanisms, not claims that every institution relies on them. Work must be analyzed before judgment.

The largest error would be drawing a straight line from “junior” to “expert.” Some experts learned mainly through instruction, difficult assignments, job rotation, or simulation—not repetition of junior tasks. AI may rescue a new worker from years of low-value work and open higher learning. This is not a marginal objection; it is the serious competing hypothesis.

An institution should not distribute work to preserve “entry-level jobs” as they were. It should preserve the **capacity-acquisition chain**, even if every tool and task changes. Routine report writing may vanish, while interpreting evidence, challenging assumptions, and confronting unfamiliar cases still have to be learned somewhere.

The expert manager’s role then changes from reviewing hundreds of pages at the end to designing a sequence in which beginners bear gradually increasing intellectual responsibility. Expert explanation and correction may cost more immediately than a machine answer produced in seconds, but it may invest in an asset absent from the balance sheet.

The Kingdom Need Not Choose Between Automation and Training

If a Saudi software, banking, or industrial company deploys an assistant at scale, it need not begin by assuming a local skills crisis; that requires local evidence. It should test whether work allocation gives new entrants a route to independent expertise.

The answer is not an “AI-free day” for everyone. That may be possible in some software exercises and dangerous or professionally impermissible in medical or operational procedures. A more suitable device is a **practice budget**: safe cases, explicit learning goals, qualified supervisors, and periodic independence tests, treated as human-capital investment rather than punishment for tool use.

In a technology company, a trainee can explain a fault before seeing an assistant’s proposal, then compare the hypothesis with an actual experiment. In finance, trainees can use privacy-cleared cases containing conflicting signals, recording supporting and disconfirming reasons before expert review. In maintenance engineering, a simulator can present controlled faults without turning real industrial machinery into an expensive classroom.

Counting hours is insufficient. A hundred hours clicking the right answer do not equal one hour explaining why a convincing hypothesis failed. What matters is changed performance on unseen problems, without assistance and with supervision appropriate to the profession’s risk.

An Institutional Experiment That Could Falsify This Article

A manager's feeling that employees are not learning does not establish the risk in one institution. Subject to professional and ethical requirements, a three- or six-month controlled experiment could compare three safe ways of working.

The first group uses an assistant to produce the solution directly under institutional policy. The second uses one that starts with questions and hints, delays the answer, and includes expert review. The third diagnoses simulated cases without direct assistance, then uses it for comparison and explanation.

"How much did each group complete?" cannot be the only measure. Record time, output quality, supervision burden, and error cost, then periodically test new problems without assistance. Also test delayed retention, detection of a deliberate error in a synthetic output, and transfer from a practiced problem to one with a different structure. Prior experience must be recorded so natural differences are not attributed to the tool.

If the first group maintains high independent-test quality, the hypothesis weakens. If the second gains independence at an acceptable cost, assistance design is preferable to prohibition. If the third excels at a skill the institution no longer needs, that is no argument against progress.

An experiment engineered to vindicate a preferred idea has no scientific value. Good institutional design allows a fair experiment to embarrass the board, not merely a presentation to persuade it.

Who Is Responsible for Producing Someone Capable of Objecting?

In some professions, years teach people to notice signals invisible to a non-specialist: a lawyer sees that sound wording hides an exception; an engineer notices a sensor reading inconsistent with the rest of the system; a doctor connects a vague symptom to a history absent from the first image. These are not magic powers or always the product of spontaneous practice. They often accumulate through knowledge, training, feedback, and graduated responsibility.

AI can accelerate that accumulation too. It can generate cases, offer competing explanations, and give trainees more attempts than human-supervisor hours permit. But excessive delivery of answers may leave an employee able to approve a result without diagnosing the error hidden inside it.

The strategic question is not whether companies will remove tasks from the first rung of a career. Some will probably disappear or change radically. It is whether a new ladder will be built, or whether the company will buy a service that carries it upstairs and later discover that nobody remembers how stairs are made.

Return to the FAA's 2013 page. It called for a mature balance: use autopilot and preserve human practice where safety permits—neither fly with automation less nor train humans at passengers' expense.

In this sense, the gravest loss in some AI programs may not be a mistake the machine made. It may be **a mistake a trainee never had the chance to make and understand**. The value of the errors we saved will emerge years later, when we need a person capable of telling an accomplished system: not this time.

References

1. SAFO 13002: Manual Flight Operations. US Federal Aviation Administration, January 4, 2013; **primary official alert**. Establishes the opening event and recommendations to preserve manual-practice opportunities while recognizing automation's gains.
2. SAFO 17007: Manual Flight Operations Proficiency. US Federal Aviation Administration, May 4, 2017; **official guidance**. Supports developing and maintaining manual proficiency in training and operational policy, subject to safety conditions.
3. The Production of Human Capital and the Life Cycle of Earnings. Yoram Ben-Porath, *Journal of Political Economy*, August 1, 1967; **original peer-reviewed economics research**. Grounds skill acquisition as investment with future returns; the article's illustrative equation is not taken from it.
4. How AI Impacts Skill Formation. Judy Hanwen Shen and Alex Tamkin, draft submitted January 28 and revised February 1, 2026; **unreviewed preliminary experimental research**. A 52-participant experiment, two groups of 26, measuring learning after practice with Trio; does not establish long-term effects in workers.
5. Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. Henry Roediger and Jeffrey Karpicke, *Psychological Science*, March 2006; **peer-reviewed experiment**. Measures retrieval's benefit for delayed retention under specified conditions; does not test AI assistants.
6. Productive Failure. Manu Kapur, *Cognition and Instruction*, July 2008; **peer-reviewed educational research**. Explains why an unsuccessful first attempt may be part of learning design when followed by suitable consolidation; not permission for dangerous workplace errors.
7. Ironies of Automation. Lisanne Bainbridge, *Automatica* 19(6), 1983; **original scientific paper**. The theoretical source of the irony in leaving unusual cases to humans after removing their ordinary practice.
8. The Retention of Manual Flying Skills in the Automated Cockpit. Stephen Casner et al., *Human Factors*, December 2014; **peer-reviewed simulation study**. A counterexample to generalizing that every flying skill declines; 16 pilots and the limits of a 747-400 simulation.

9. Endoscopist Deskilling Risk after Exposure to Artificial Intelligence in Colonoscopy: A Multicentre, Observational Study. Krzysztof Budzyń et al., *The Lancet Gastroenterology & Hepatology*, online August 12, 2025, October issue; **multicenter observational study**. 1,443 unassisted examinations; 28.4 versus 22.4 percent. Does not isolate all confounding or establish causality.
10. Artificial Intelligence Augmented Tutoring vs Expert Instruction on Learning Simulated General Surgical Skills: A Systematic Review and Meta-analysis. Fuad Hanna et al., *BMC Medical Education*, June 10, 2026; **peer-reviewed systematic review and meta-analysis**. Four studies and 268 participants; suggests augmented teaching may work, with low certainty, not that human supervisors can be removed.
11. Generative AI without Guardrails Can Harm Learning: Evidence from High School Mathematics. Hamsa Bastani, Osbert Bastani, Alp Sungu et al., *PNAS*, online June 25, July 1 issue, 2025; **peer-reviewed randomized field experiment**. Compares practice performance with an independent test among nearly one thousand Turkish students and distinguishes general from hint-constrained assistance.
12. Does Using Artificial Intelligence Assistance Accelerate Skill Decay and Hinder Skill Development without Performers' Awareness?. Brooke Macnamara et al., *Cognitive Research: Principles and Implications*, July 12, 2024; **hypothesis-setting research analysis**. Useful for defining questions, not an experiment proving skill loss.
13. Operational Use of Flight Path Management Systems: Final Report. US safety agencies' Flight Deck Automation Working Group, September 8, 2013; **official technical report**. Documents integration of training and system operations in safety assessment, not an experiment on language models.
14. The Impact of AI on Developer Productivity: Evidence from GitHub Copilot. Sida Peng, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer, draft dated February 13, 2023; **controlled experiment in a preliminary paper**. Shows a possible 55.8 percent acceleration on a programming task in context, without directly measuring long-term skill formation.