

Deskilling Is Not One Endpoint: Five Quantities Behind Assisted Gains and Unassisted Losses in Clinical AI, Why Two Studies of the Same Design Disagree, and Why a Within-Clinician Trial's Control Arm Has Already Been Exposed

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. Clinical AI tools are approved and bought on one endpoint, the performance of a clinician working with the tool, while a second literature asks what happens to the same clinician working without it. The two are usually set against each other as gain against loss. This paper argues that the comparison is malformed, because "deskilling" names at least five different quantities, the studies most often cited measure different ones, and the trials that establish the gain are themselves exposed to the loss. The five are the assisted effect against an unexposed clinician, unassisted performance while the aid is in routine use on other cases, unassisted performance after the aid has been withdrawn for a period, performance when the aid is present but wrong, and the skill a learner acquires when trained with the aid. The colonoscopy evidence illustrates the problem. A multicentre observational study reported a fall in adenoma detection on non-AI procedures from 28.4% to 22.4% after AI was introduced; a prospective three-phase trial found no significant change after removal; and a single-centre study with AI in half the rooms found non-AI detection maintained. The first and third measure the same quantity and disagree, so the conflict is not only a confusion of estimands. The paper then generalises a point the authors of the Polish study made about their own data, where AI-assisted procedures in the post-AI period reached 25.3%, above the concurrent non-AI 22.4% but below the pre-AI 28.4%: when cases are randomised within clinician, the control arm is unassisted work by exposed clinicians, so any spillover from exposure, in either direction, biases the trial estimate of the assisted gain. Facility-level cluster randomisation avoids this. Neither literature reports the share of cases handled without the aid, which is the term that converts any deskilling estimate into a deployment cost. The paper consolidates the published values, states a decision procedure that makes each missing term explicit, and specifies three designs that would settle the open parts.

Keywords: deskilling; clinical artificial intelligence; computer-aided detection; colonoscopy; mammography screening; automation bias; skill retention; study design

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

The argument for putting an AI aid in front of a clinician is usually a comparison between two arms of a trial: procedures done with the aid and procedures done without it. In colonoscopy, randomised trials of computer-aided detection (CADE) report higher adenoma detection rates with the aid, and a meta-analysis of the first five such trials pooled 36.6% against 25.2% (Hassan et al., 2021). In mammography screening, the MASAI trial in Sweden found that AI-supported reading detected more cancers, cut the screen-reading workload by 44.2%, and was non-inferior on interval cancers (Hernstrom et al., 2025; Gommers et al., 2026). These are the numbers on which guideline panels deliberate and purchasers act.

A second literature asks a question those trials were not built to answer: what happens to a clinician's own performance once the aid is part of daily work. Its most discussed result comes from four Polish endoscopy centres, where adenoma detection on colonoscopies performed without AI fell from 28.4% in the three months before AI was introduced to 22.4% in the three months after (Budzyn et al., 2025). The finding drew an

editorial comment (Ahmad, 2025), several letters (Lam, 2025; Levartovsky and Kopylov, 2025; van de Sande et al., 2025), an authors' reply (Budzyn, Romanczyk and Mori, 2025) and wide coverage, and it sits alongside reviews and commentaries arguing that deskilling is a general risk of clinical AI (Natali et al., 2025; Heudel et al., 2026; Topaz et al., 2026). The question has reached regulatory commentary as well, which asks whether deskilling by a medical AI system is prohibited or permitted under the EU AI Act (Gerke, Hassan and Mori, 2026).

The usual framing sets the two literatures against each other: gains in assisted performance on one side, losses in unassisted performance on the other, with deployment decisions supposed to weigh them. This paper argues that the framing cannot support the weighing it is asked to do, for three reasons, each of which is developed in a section of its own.

The first reason is that "deskilling" is not one quantity. The published studies measure at least five different things: the assisted effect relative to a clinician who has never used the aid; unassisted performance while the aid is in routine use on other cases; unassisted performance after the aid has been withdrawn for a period; performance on cases where the aid is present but

wrong; and the skill acquired by a learner who trains with the aid. These quantities have different causes, different time courses and different deployment consequences, and a result about one does not transfer to another (Sections 3 to 8).

The second reason is that the conflict is not only a confusion of quantities. Two colonoscopy studies that measure the same thing, unassisted performance while the aid is in routine use, report opposite findings: the Polish study a fall, and a Japanese single-centre study with CADe installed in half the rooms no decline (Budzyn et al., 2025; Okumura et al., 2026). A customer-support study that measured the same quantity during software outages found unassisted performance above the pre-AI baseline (Brynjolfsson, Li and Raymond, 2025). Separating the quantities narrows the disagreement; it does not dissolve it (Section 5).

The third reason is structural, and for colonoscopy it has already been stated by the authors of the Polish study, who argue that a reduced ADR on non-AI procedures "might help create significant differences of ADR" in randomised trials of the aid (Budzyn et al., 2025). This paper generalises their point. Many studies of clinical aids randomise or alternate cases, days or rooms while the same clinicians work in both conditions; the centres in the most discussed deskilling study assigned colonoscopies to AI or no AI by examination date (Budzyn et al., 2025). If exposure to the aid changes how clinicians perform without it, in either direction, then the control arm of such a trial is not an unexposed comparison: it is unassisted work by exposed clinicians. The trial then estimates the assisted gain against a moved baseline. If the Polish result is correct, the assisted gain reported by within-clinician trials is overstated relative to practice without the aid; if the outage result generalises, it is understated. The two literatures are not independent bodies of evidence to be weighed; one describes the comparison arm of the other (Section 9).

A fourth observation follows from the first three. The quantity that converts any deskilling estimate into a deployment cost is the share of cases a clinician handles without the aid, through outages, out-of-scope cases, workflow gaps or voluntary non-use. No study in either literature located for this review reports it. Until it is measured, a deskilling result, whatever its sign, is a statement about a counterfactual rather than an estimate of harm (Sections 10 and 11).

The contribution is therefore not a verdict on whether clinical AI deskills. It is a decomposition of the claim into quantities that can each be estimated, an audit of which designs estimate which, a generalisation of the Polish study's argument that one widely used trial design entangles the gain with the loss, and a decision structure that makes the unmeasured terms explicit. Section 2 states scope and terms. Sections 3 to 8 work through the five quantities. Section 9 develops the control-arm argument. Section 10 consolidates the numbers in Table 1 and Figure 1. Section 11 gives Algorithm 1. Section 12

takes the strongest objections. Section 13 proposes designs. Sections 14 to 16 state what is not known, the paper's limitations, and its conclusion.

2. Scope, terms and how the evidence was read

2.1 What the paper covers

The paper covers decision aids that act on a clinician's perception or judgement during a task the clinician could, in principle, perform alone: polyp detection during colonoscopy, lesion detection and triage in mammography, image classification in radiology and dermatology, and diagnostic suggestion in hospital medicine. Evidence from aviation, education, programming and customer support is used where it measures one of the five quantities more cleanly than the clinical literature does, and its transfer to clinical settings is flagged each time.

2.2 What it does not cover

It does not cover autonomous systems that replace the clinician on a task, where unassisted performance is no longer a deployment variable. It does not assess whether CADe or AI-supported screening should be adopted; that question depends on patient-level outcomes, overdiagnosis and cost, which two guideline panels have weighed and reached different conclusions on (Sultan et al., 2025; Foroutan et al., 2025). It does not attempt a systematic review; the deskilling literature has recently been reviewed several times (Natali et al., 2025; Heudel et al., 2026), and those reviews are cited for what they establish, not repeated.

2.3 Terms

An *aid* is any AI output presented to a clinician during a task: a bounding box on a colonoscopy frame, a malignancy score and marks on a mammogram, a suggested category, a probability. *Assisted performance* is the clinician's performance on cases where the aid is presented. *Unassisted performance* is the clinician's performance on cases where it is not. *Exposure* is the clinician's history of assisted work. A *within-clinician* design randomises or alternates cases, days or rooms while the same clinicians perform both arms; a *cluster* design assigns the aid to facilities or clinicians, so that the comparison arm consists of clinicians who are not exposed.

Detection metrics are used as their sources define them. The adenoma detection rate (ADR) is the proportion of colonoscopies with at least one histologically confirmed adenoma. PDR-5, used by the Norwegian registry, is the proportion of colonoscopies with at least one polyp of 5 mm or larger, which the registry treats as a surrogate for ADR (Pedersen et al., 2026). These are not interchangeable, and Table 1 keeps them apart.

2.4 How the evidence was read

Every source was located through a registry search and verified against its Crossref, PubMed or arXiv record. Full texts were read where they were openly available

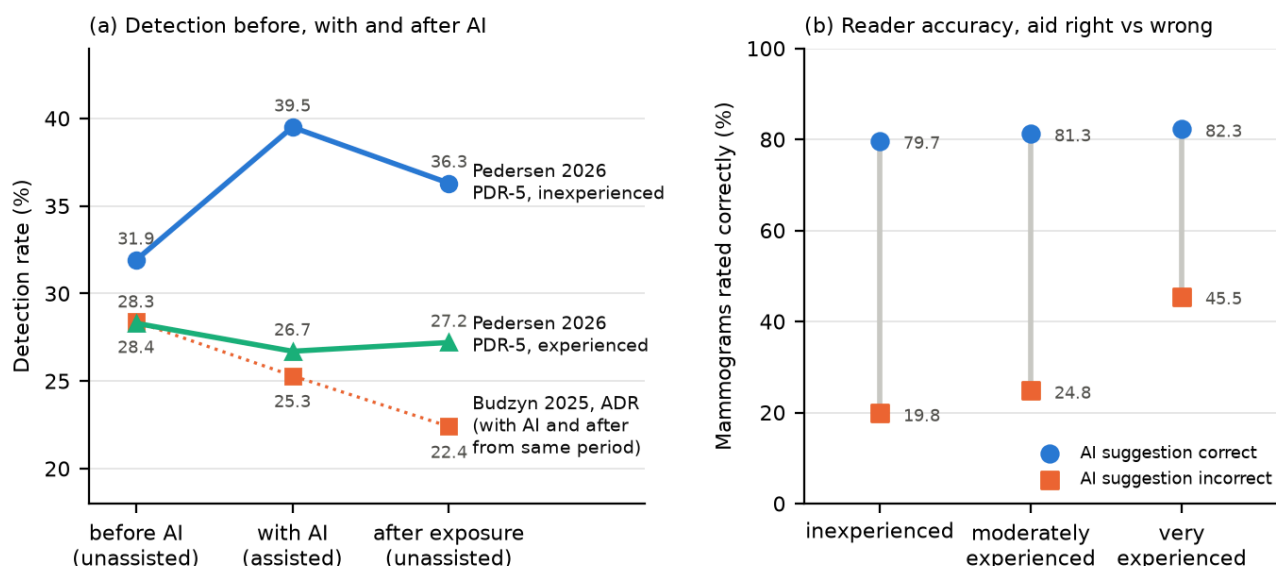


Figure 1. Two of the quantities in published numbers. Panel (a): detection on colonoscopies before AI exposure, with AI, and on unassisted colonoscopies after exposure. Budzyn et al. (2025) report ADR on non-AI procedures in the three months before and after AI was introduced, with AI in use on other days in the second period (Q2), and ADR on the AI-assisted procedures of that second period (25.3%); its "with AI" and "after exposure" points come from the same three months. Pedersen et al. (2026) report PDR-5 for inexperienced and experienced endoscopists before CAde, with CAde, and after CAde was removed (Q1 and Q3). ADR and PDR-5 are different metrics and the two studies' levels are not comparable. Panel (b): percentage of mammograms rated correctly by 27 radiologists when a purported AI suggested the correct versus an incorrect BI-RADS category, by experience (Dratsch et al., 2023; Q4). Each panel re-plots values reported by the cited sources, and every plotted number is printed there.

and the paper relies on more than the abstract: the three-phase colonoscopy trial (Pedersen et al., 2026), the customer-support study (Brynjolfsson, Li and Raymond, 2025, read in its arXiv version), two randomised experiments on learning with AI (Liu et al., 2026; Shen and Tamkin, 2026) and the skill-decay meta-analysis (Arthur et al., 1998). The Polish observational study (Budzyn et al., 2025) was read in its openly available accepted manuscript (Warwick Research Archive Portal, wrap.warwick.ac.uk/191005). The letters and reply that followed it were not accessible in full, and the paper does not characterise their content. This is a real limitation, because the correspondence is where the study's design was most directly contested, and Section 15 returns to it.

3. Five quantities called deskilling

The claim that AI "deskills" clinicians is used for at least five different comparisons. Each is defined here as a contrast between two performance levels. The definitions are stipulations, and the rest of the paper uses them.

Q1, the assisted effect. Assisted performance minus the unassisted performance of a clinician who has not been exposed to the aid. This is the quantity regulators and purchasers want. It is what a trial estimates when its comparison arm is genuinely unexposed.

Q2, the interleaved-withdrawal effect. Unassisted performance of a clinician who is using the aid routinely on other cases, minus that clinician's unassisted performance before exposure. The unassisted cases are interleaved with assisted ones: a non-AI day between AI

days, a room without the device down the corridor from one with it, an outage in the middle of a shift.

Q3, the sustained-withdrawal effect. Unassisted performance after the aid has been withdrawn for a period, minus pre-exposure unassisted performance. This is the retention question: whether exposure has changed what the clinician can do when the aid is reliably absent. The distinction between Q2 and Q3 is drawn explicitly by Pedersen et al. (2026), who contrast their own design, which measured exposure "'after' CAde was removed", with the Polish study, which measured non-CAde work "'while' they were exposed to CAde in the other colonoscopies they performed during the same time period".

Q4, the failure-mode effect. Assisted performance on cases where the aid is wrong, minus unassisted performance on the same cases. This is automation bias in its clinical form, and it is the term that determines how much of the aid's error passes through the clinician.

Q5, the formation effect. The unassisted performance of a learner trained with the aid, minus that of a learner trained without it, tested once training ends. A recent review of educational strategies for supervising AI use in clinical training separates this from deskilling under the name never-skilling, alongside mis-skilling, the uptake of the aid's errors (Abdulnour, Gin and Boscardin, 2025).

These quantities differ in mechanism. Q2 can move without any change in capacity: a clinician who has learned to expect a box around every polyp may search less carefully on a day without boxes, and recover as soon as the box is reliably gone. Q3 isolates a change in

capacity from such a change in expectation. Q4 can be large even when Q2 and Q3 are zero, because it measures what happens when the aid's output is followed, not what happens when it is absent. Q5 concerns people who have not yet acquired the skill, so a null result for experts says nothing about it.

They also differ in deployment consequence. Q1 enters the expected value of every assisted case. Q2 matters in proportion to the share of cases handled without the aid during routine use. Q3 matters if the aid may be withdrawn, through a procurement decision, a regulatory action or a device failure lasting more than a shift. Q4 matters in proportion to the aid's error rate on the cases it sees. Q5 matters to the next cohort of clinicians, whose unassisted baseline is set by how they were trained.

The human-factors literature anticipated this structure. Bainbridge (1983) argued that automating the routine part of a task leaves the operator with the cases automation cannot handle, while removing the practice that would have kept them able to handle those cases. Endsley and Kiris (1995) linked out-of-the-loop performance after an automation failure to reduced situation awareness, which they attributed mainly to the shift from active to passive processing. A meta-analysis of 18 experiments found that a higher degree of automation improved routine performance and reduced workload but impaired performance when the automation failed, with the sharpest costs where automation moved from supporting information analysis to supporting action selection (Onnasch et al., 2014). In the vocabulary above, that result says Q1 and Q4 trade off. What the clinical literature adds is evidence on Q2, Q3 and Q5, and a design problem that the laboratory literature did not face.

4. The assisted endpoint, and what it can and cannot register

4.1 What the trials measure

The CADe trials measure Q1 under the conditions of a trial. In a three-centre Italian trial of 685 patients, ADR was 54.8% with CADe and 40.4% without, a relative risk of 1.30, with no significant difference in withdrawal time (Repici et al., 2020). The early meta-analysis pooled five trials and 4,354 patients (Hassan et al., 2021). By the time the American Gastroenterological Association panel deliberated, it accepted an 8-percentage-point increase in ADR (95% CI 6 to 10) and a 2-point increase in detection of advanced adenomas or sessile serrated lesions (Sultan et al., 2025).

That gain has not held up consistently outside trials. A meta-analysis of eight non-randomised studies with 9,782 patients found ADR of 44% with CADe and 38% without, a relative risk of 1.11 whose interval included no effect (Patel et al., 2024). Several pragmatic trials were null. In four community centres in the United States, adenomas per colonoscopy were 0.73 with CADe and 0.67 without, and ADR was 35.9% against 37.2%

(Wei et al., 2023). In an international trial in which experienced endoscopists performed every procedure, ADR was 38.4% against 37.7%, though the authors note that the control-arm ADR exceeded their sample-size assumption, which raises the risk that the trial was underpowered (Maas et al., 2024). A pragmatic implementation trial at a US academic centre reported no improvement in colonoscopist performance (Ladabaum et al., 2023), and a survey and interviews of its colonoscopists afterwards found high trust in AI and no unifying explanation for the null result (Watkins et al., 2024). Against these, a cluster-randomised quality-improvement study across Veterans Health Administration facilities found that ADR rose by 4.2 percentage points at facilities given CADe and fell by 0.7 points at control facilities, an adjusted odds ratio of 1.22 (Dominitz et al., 2026). Reviews of the gap between efficacy and effectiveness list the absence in routine practice of the Hawthorne effect that trial participation produces, the quality of mucosal exposure and decisions about diminutive polyps as candidate explanations (Putera, Koh and Li, 2026), and deskilling appears in the same reviews as a harm still to be studied.

4.2 What the guidelines concluded

Two guideline panels read essentially the same trial evidence in 2025 and reached different conclusions. The AGA made no recommendation for or against CADe, citing very low certainty about cancer outcomes, a modelled 11 fewer colorectal cancers and 2 fewer deaths per 10,000 people, and 635 more surveillance colonoscopies per 10,000 (Sultan et al., 2025). A panel convened with MAGIC and The BMJ, drawing on 44 randomised trials with more than 30,000 participants and a microsimulation, suggested against routine use (Foroutan et al., 2025). Neither guideline's abstract lists deskilling among its reasons; both centre on whether a detection gain concentrated outside advanced lesions translates into fewer cancers. The European society's response has been a curriculum rather than a recommendation, one of whose six elements is continuous monitoring of performance indicators after AI is introduced (Mori et al., 2026).

4.3 The mammography case

Mammography offers a cleaner Q1 estimate and a historical warning. MASAI randomised more than 105,000 women in Sweden between AI-supported screening and standard double reading. The AI score triaged examinations to single reading (scores 1 to 9) or double reading (score 10), and radiologists saw the scores and, for higher-risk examinations, detection marks (Lang et al., 2023). AI-supported screening detected 6.4 cancers per 1,000 against 5.0 (Hernstrom et al., 2025), and the interval cancer rate was 1.55 per 1,000 against 1.76, a ratio of 0.88 that met the non-inferiority criterion, with sensitivity of 80.5% against 73.8% and identical specificity (Gommers et al., 2026). A German observational implementation study across 463,094 women found a 17.6% higher detection rate where

radiologists chose to use AI support (Eisemann et al., 2025).

The warning comes from the previous generation of computer-aided detection in mammography. It was approved in the United States in 1998 and spread rapidly. An analysis of 43 facilities found that the seven that implemented it saw specificity fall from 90.2% to 87.2% and biopsies rise by 19.7%, with no significant change in sensitivity or cancer detection (Fenton et al., 2007). A larger analysis of 271 radiologists from 2003 to 2009 found no improvement on any metric. Among the 107 radiologists who read both with and without the tool, sensitivity was lower with it, an odds ratio of 0.53 (Lehman et al., 2015). The same analysis notes that approval in 1998 and higher Medicare payment from 2002 preceded any established benefit. A reanalysis of a reader study that had detected no significant average effect found that the null concealed a split: the tool raised sensitivity slightly for the 44 least discriminating readers on easy cancers and lowered it by 0.145 for the six most discriminating readers on difficult ones (Povyakalo et al., 2013).

4.4 What the assisted endpoint cannot register

Three things follow for the deskilling question. First, a trial's assisted arm measures performance when the aid is present, so by construction it cannot register Q2, Q3 or Q5, all of which concern performance when the aid is absent. Second, it registers Q4 only as an average over cases where the aid is right and cases where it is wrong, weighted by the aid's error rate in the trial population; if the error rate differs in deployment, the average does not transfer. Third, and least appreciated, its comparison arm registers Q2 whenever the same clinicians perform both arms, which is the subject of Section 9. MASAI itself randomised examinations, not radiologists, and the radiologists were not masked to group (Lang et al., 2023). The abstracts read for this review do not state whether the same radiologists read both arms, and the full texts could not be retrieved for this paper, so the point is left open and the extension of Section 9 to MASAI is conditional on it. If they did, as the design permits, its control readings were done by radiologists who were reading AI-supported examinations in the same period.

MASAI also changes something that mammography research has linked to performance: how much reading each radiologist does. AI triage cut screen readings from 109,692 to 61,248 (Hernstrom et al., 2025). Whether lower reading volume affects accuracy is itself contested. In one US consortium analysis, sensitivity did not differ by total, screening or diagnostic volume, while false-positive rates were higher for the lowest-volume radiologists (Buist et al., 2011). The volume channel is therefore a hypothesis about Q3 in screening, not evidence for it, and MASAI's two-year interval-cancer endpoint is too short to test it.

5. Interleaved withdrawal: the same quantity, opposite answers

5.1 The Polish study

The study by Budzyn et al. (2025) is a Q2 study. It was conducted at four centres taking part in the ACCEPT trial, which brought AI polyp-detection tools into use in late 2021; from then on, colonoscopies were assigned to be done with or without AI according to the examination date. The study compared non-AI colonoscopies in the three months before AI was introduced (795 patients) with non-AI colonoscopies in the three months after (648 patients). The 19 endoscopists each had performed more than 2,000 colonoscopies. ADR fell from 28.4% to 22.4%, an absolute difference of -6.0 percentage points (95% CI -10.5 to -1.6). In multivariable analysis, exposure to AI had an odds ratio of 0.69, alongside patient sex and age. On the 734 AI-assisted colonoscopies of the post-AI period, ADR was 25.3%; in a supplementary model comparing them with the pre-AI procedures, the authors report that "use of AI did not significantly affect the ADR". The abstract concludes that continuous exposure "might reduce" ADR on non-AI procedures, "suggesting a negative effect on endoscopist behaviour".

The wording is careful, and it matters. A before-after comparison over six months with no concurrent unexposed group cannot separate an effect of exposure from anything else that changed between the two windows: the case mix, the referral pattern, staffing, or the act of joining a trial. The authors address part of this themselves. They point out that the post-AI non-AI procedures had "more male patients (38.0% vs. 45.4%, $P=0.005$) and greater sedation use (77.1% vs. 81.9%, $P=0.024$)—both factors that would typically increase ADR", so these two imbalances work against the observed fall. They also report that the post-AI phase nearly doubled in procedure volume, because Poland's colonoscopy screening programme was suspended from January 2022 and the centres took more diagnostic patients from their waiting lists, which they list as a potential confounder. And they note that the large drop in the two centres with the highest baseline ADR "could be the result of regression to the mean". The study's population was diagnostic colonoscopies rather than a screening programme, so its baseline ADR of 28.4% is not directly comparable with screening-trial figures. And the post-period non-AI procedures were interleaved with AI procedures by date, so the result estimates Q2, not Q3: it describes what endoscopists did on non-AI days while AI was in use on other days, not what they could do after it was gone.

5.2 The Japanese study

Okumura et al. (2026) also estimated Q2, with a different design. At a single centre from 2021 to 2023, CAdE systems were installed in half the endoscopy rooms and patients were distributed equally between rooms. After propensity-score matching, 4,712 procedures were analysed, half with CAdE and half

without. Endoscopists were classified by their ADR before CADE as high detectors (25% or more) or low detectors. Using cumulative-sum learning curves, the study found that both groups improved on CADE procedures; on non-CADE procedures, high detectors' learning curves accelerated and low detectors' did not change. The study concludes that, once CADE was in place, detection on the procedures done without it held steady over the study period rather than falling.

Both studies estimate unassisted performance while the aid is in routine use on other cases. Both have endoscopists alternating between assisted and unassisted procedures. They point in opposite directions. Okumura's design has a concurrent comparison, which Budzyn's does not, but it spans three years, so slow trends in endoscopist skill are mixed into the curves, a concern Pedersen et al. (2026) raise explicitly. Budzyn's design spans six months, but without a concurrent comparison. Neither is a randomised estimate of Q2.

5.3 The outage study

Brynjolfsson, Li and Raymond (2025) studied the staggered introduction of a generative-AI assistant among 5,172 customer-support agents. Access raised issues resolved per hour by 15% on average, with the largest gains for less experienced and lower-skilled agents. Their analysis of software outages, periods when the assistant gave no suggestions, is an estimate of Q2. In the arXiv version of the paper, which this review read, the authors report that agents who had been exposed to the assistant handled chats faster than their pre-AI baseline even during outages, that this outage-period gain grew with time since adoption, and that it appeared among agents who had initially followed the suggestions closely and not among those who frequently deviated from them. The authors caution that outages are rare and not random, and that the outage estimates are noisy.

5.4 What could explain the disagreement

Three Q2 studies, then: one negative, one null-to-positive, one positive. The disagreement is real, and this paper does not resolve it. It is possible to say what would have to differ for all three to be correct.

One candidate is what the aid's output contains. A conversational suggestion carries content: a phrasing, a procedure, a fact about the product. An agent who reads and uses it may retain that content when the suggestions stop. A CADE bounding box carries a location. It tells the endoscopist where to look on this frame, and nothing that transfers to the next patient. If the aid supplies content, exposure can teach; if it supplies a pointer, exposure can only redistribute attention. This paper suspects that the difference between conversational and perceptual aids accounts for part of the sign difference between the outage study and the Polish study, but no study located for this review compares the two kinds of aid within one design, so the suggestion is a conjecture.

A second candidate is how the clinician engages with the aid. In the customer-support data, the outage gains

appeared among agents who had followed the suggestions closely (Brynjolfsson, Li and Raymond, 2025). In Okumura et al. (2026), the non-CADE gains appeared among high detectors and not low ones. Both point to the clinician's existing practice as a moderator. Neither study randomised engagement, so both are associations.

A third candidate separates effort from capacity. Colonoscopy has a routinely recorded proxy for effort: withdrawal time, the minutes spent inspecting the mucosa on the way out. If Q2 is negative because endoscopists search less carefully on non-AI days, withdrawal time on those days should fall; if it is negative because they have lost some perceptual skill, withdrawal time need not change. The published values show that the aid itself moves the proxy in some settings and not others. In the Italian trial, withdrawal time did not differ between arms, 417 against 435 seconds (Repici et al., 2020). In the US community trial, it was longer with CADE, 11.7 against 10.7 minutes, but similar when no polyp was found (Wei et al., 2023). In the Norwegian trial, both experience groups had longer withdrawal times during the CADE phase, although the registry records withdrawal time only for procedures without biopsy or polypectomy, so the authors treat it with caution (Pedersen et al., 2026). Withdrawal time on the non-AI procedures before and after exposure is the comparison that would discriminate between the two readings of the Polish result, but the Polish study did not record it; its authors list "the lack of data of endoscope withdrawal time" among their limitations (Budzyn et al., 2025). The discrimination therefore needs new data or a registry that records withdrawal time.

A fourth candidate is the length and structure of exposure. Budzyn's post-period was the first three months after AI arrived, which is when novelty and adjustment effects would be largest. Okumura's covered three years, and the outage gains in the customer-support data grew over months. A short-window negative result and a long-window null are compatible if Q2 is transiently negative and then recovers, though neither study measured that trajectory. Pedersen et al. (2026) offer a related transient reading when comparing their null result after removal with the Polish result, "suggesting that the deskillling effect may be a temporal phenomenon that disappears over time after CADE is removed".

6. Sustained withdrawal: what remains when the aid is gone

6.1 The Norwegian trial

Pedersen et al. (2026) designed a study of Q3. Thirteen endoscopists at three Norwegian centres, seven with between 100 and 1,000 previous colonoscopies and six with more, each performed colonoscopies in three consecutive phases: before CADE, with CADE, and after CADE was removed. Each endoscopist served as their own control, and 5,013 colonoscopies were analysed

between November 2021 and June 2025. The outcome was PDR-5.

Among inexperienced endoscopists, PDR-5 rose from 31.9% before CADe to 39.5% with it (odds ratio 1.43, 95% CI 1.11 to 1.84) and was 36.3% after removal, which did not differ significantly from baseline (odds ratio 1.03, 95% CI 0.79 to 1.34). Among experienced endoscopists, PDR-5 was 28.3%, 26.7% and 27.2% across the three phases, with no significant effect of CADe in any comparison. At the level of individual endoscopists, three of the inexperienced group declined between the first and third phases, two were unchanged and two improved; among the experienced group, three declined and three improved. The authors conclude that no upskilling or deskillng was observed.

The design isolates Q3 more cleanly than any other colonoscopy study, and its limits are stated by its authors. Some inexperienced endoscopists performed few colonoscopies in the third phase, so that comparison may be underpowered. Case mix differed between phases for the inexperienced group. And the phases were long. For inexperienced endoscopists, mean inclusion time from the first colonoscopy was 6.8 months in the first phase, 17.7 in the second and 23.8 in the third, so their ordinary learning over two years runs alongside any effect of CADe. A fixed-order before-with-after design cannot separate the two. For experienced endoscopists, the trial found no effect of CADe on PDR-5 even while it was in use, although their withdrawal times, like those of the inexperienced group, were longer during the CADe phase (Pedersen et al., 2026). A null detection result is therefore not evidence that the aid left their behaviour unchanged, and the only negative clinical Q2 estimate comes from experienced endoscopists (Budzyn et al., 2025).

6.2 What aviation found

Aviation has studied Q3 for longer and with better instruments. In a Boeing 747-400 simulator, 16 airline pilots flew routine and non-routine scenarios at different levels of automation (Casner et al., 2014). Their instrument scanning and manual control were largely intact, even for skills they reported practising rarely. When asked to perform the cognitive work of manual flight unaided (tracking position without a map display, deciding the next navigational step, recognising instrument failures), they made more frequent and more serious errors, and those errors were associated with how often pilots reported task-unrelated thought while automation was flying. A study of 126 airline pilots flying a raw-data instrument approach found long-haul pilots performed worse than short-haul pilots, with recent flight practice a stronger predictor of fine-motor performance than total experience or time since training (Haslbeck and Hoermann, 2016). An earlier study had related manual handling performance to recent flying experience in air transport pilots (Ebbatson et al., 2010).

Two lessons carry over. First, Q3 is component-specific and practice-specific. In the simulator study, manual

control largely held while the cognitive components of manual flight degraded (Casner et al., 2014); among the 126 pilots, fine-motor performance degraded where recent manual practice was scarcest (Haslbeck and Hoermann, 2016). The two results are consistent if what decays is whichever component has stopped being practised, and they show that a single aggregate outcome can miss the loss. A colonoscopy outcome that aggregates insertion, mucosal inspection, lesion recognition and the decision to resect cannot detect a loss confined to one component. Second, the predictor was recent practice of the specific component, not exposure to automation as such. The general skill-retention literature agrees. A meta-analysis of 189 independent data points found skill loss ranging from near zero immediately after training to an effect size of about -1.4 after more than a year of non-use, with cognitive and accuracy-based tasks more susceptible than physical and speed-based ones (Arthur et al., 1998). On that evidence, the relevant exposure variable for Q3 is how much unassisted practice of the specific component a clinician still gets, which no colonoscopy study reports.

6.3 What controlled experiments found

Outside medicine, randomised experiments have estimated short-horizon versions of Q3 directly. Across experiments with 1,222 participants solving fraction and reading-comprehension problems, about ten minutes of AI-assisted work was followed by worse unassisted performance and more giving up than in a control group (Liu et al., 2026). In the replication experiment, the solve rate on the final three unassisted problems was 0.71 for the AI group against 0.77 for controls, a Cohen's *d* of -0.19; within the AI condition, participants who had used the AI to obtain answers directly fell to 0.65, while those who had used it for hints scored 0.76. In a field experiment with nearly a thousand high-school mathematics students, unrestricted access to GPT-4 raised grades by 48% while available and lowered them by 17% relative to never-access students once removed; a tutor version with safeguards raised grades by 127% while available and largely avoided the loss (Bastani et al., 2025).

These experiments establish that Q3 can be negative, can appear quickly, and depends on how the aid is used. They do not establish that it is negative for expert clinicians performing a well-practised task. Their participants were learning or rehearsing, their horizons were minutes to weeks, and their aids supplied answers rather than pointers. They are closer to Q5 than to the experienced-endoscopist case.

7. Failure mode: performance when the aid is wrong

Q4 is the best-measured of the five in controlled settings and the one least often called deskillng, though in deployment it may carry the largest cost. It is estimated by presenting clinicians with cases where the aid is

deliberately wrong.

In a mammography experiment, 27 radiologists read 40 test mammograms with a purported AI system that suggested an incorrect BI-RADS category for 12 of them (Dratsch et al., 2023). The proportion rated correctly fell from 79.7% to 19.8% for inexperienced readers when the suggestion was wrong, from 81.3% to 24.8% for moderately experienced readers, and from 82.3% to 45.5% for very experienced readers (Figure 1b). Experience attenuated the effect but did not remove it. The comparison is between different mammograms, those with a correct suggestion and those with an incorrect one, read with the aid in both cases; it is therefore a proxy for Q4 as defined in Section 3, not an estimate of assisted against unassisted reading of the same cases. In a randomised vignette study of 457 hospital clinicians, a standard AI model raised diagnostic accuracy by 2.9 percentage points over a 73.0% baseline, and by 4.4 points with explanations; a systematically biased model lowered it by 11.3 points, and explanations did not significantly reduce that harm (Jabbour et al., 2023). In a dermatology study, AI support of good quality improved accuracy beyond either the clinicians or the AI alone and helped the least experienced most, while faulty AI misled clinicians across the experience range, experts included (Tschandl et al., 2020). In a chest-radiograph experiment, inaccurate advice degraded diagnostic accuracy whether it was labelled as coming from an AI or from a human expert (Gaube et al., 2021).

Two reviews place these results in a longer literature. One synthesised automation bias in clinical decision support and catalogued its frequency, mediators and mitigators (Goddard, Roudsari and Wyatt, 2012). Another, covering 40 studies across domains, six of them in healthcare, found automation bias in single tasks, typically diagnostic ones whose output is hard to verify, rather than only under the multitasking conditions where the human-factors literature had located it (Lyell and Coiera, 2017). An earlier integration argued that complacency and automation bias share an attentional basis, reported that complacency is not overcome by simple practice, and reported that automation bias is not prevented by training or instructions (Parasuraman and Manzey, 2010).

Q4 also helps explain the heterogeneity in Q1. Among 140 radiologists across 15 chest-radiograph tasks, the effect of AI assistance varied widely between individuals. Experience, subspecialty and familiarity with AI tools did not reliably predict who benefited, lower-performing radiologists did not consistently gain more, and AI errors drove much of the harm (Yu et al., 2024). A controlled experiment found that AI helps most for people with low baseline ability who are calibrated about it (Caplin et al., 2026), and a replication on radiologists from a public chest-radiograph data set, 68 radiologists and 11,420 paired observations, reached the same conclusion (Martin, 2026; Agarwal et al., 2023). Calibration matters because the clinician must decide, case by case, whether to trust their own reading or the aid's, and that decision

is exactly where Q4 is generated.

The connection to deskilling is direct. If exposure reduces a clinician's independent reading of the case, through reduced vigilance or reduced practice, then it raises the probability that an aid error passes through uncorrected. Q4 would then grow with exposure even if Q2 and Q3, measured on cases without the aid, stayed at zero. No study located for this review measures Q4 at two levels of exposure in the same clinicians. That is the most direct test of whether deskilling, in the sense that matters for patient safety, occurs at all.

8. Formation: the next cohort's baseline

Q5 concerns learners. Its clinical evidence is thin and its non-clinical evidence is mixed in sign.

The negative evidence is experimental. Developers learning a new asynchronous programming library with AI assistance scored 17% lower on a subsequent quiz than those who learned without it, a difference of 4.15 points on a 27-point quiz (Cohen's *d* of 0.738), and did not finish significantly faster; three of six interaction patterns identified in screen recordings preserved learning (Shen and Tamkin, 2026). In a logic-puzzle experiment with phases before, during and after AI access, participants who requested assistance performed worse after it was removed, and their later unassisted performance was overestimated when predicted from their assisted performance (Wu et al., 2026). That last finding bears directly on Section 4: assisted performance is not merely silent about unassisted performance, it can mislead about it. A small electroencephalography study reported weaker connectivity in participants who wrote essays with a language model and continued weak engagement when they were switched to writing unaided, though only 18 of its 54 participants completed the crossover session (Kosmyna et al., 2025). A cognitive-science perspective argued that AI assistance may accelerate decay in experts and impede acquisition in learners without either group noticing (Macnamara et al., 2024).

The positive evidence is observational. The diffusion of chess computers helped players improve by standing in for scarce human training partners, though players who trained with machines did not learn to exploit characteristically human mistakes (Gaessler and Piezunka, 2023). Across 5.8 million move decisions by professional Go players from 1950 to 2021, decision quality rose after superhuman AI programs appeared, together with a rise in novel moves (Shin et al., 2023). In competitive programming, a study of practice histories found that entrants of the AI era showed no erosion of skill in proctored contests where AI is prohibited, and argued that substituting AI for practice and using it to learn faster look alike in AI-aided output and differ only under unaided testing (Yao, 2026).

The clinical evidence on Q5 is one subgroup. Pedersen et al. (2026) found that inexperienced endoscopists' PDR-5 rose with CADE and did not differ significantly from

baseline after removal, with the individual endoscopists split in both directions. That is not a formation study, since the endoscopists already had at least 100 procedures and the outcome was measured over two years in which ordinary learning was also at work.

8.1 Who gains most, and who is most exposed

One pattern recurs across the five quantities, and it compounds rather than offsets. The clinicians who gain most from an aid under Q1 tend to be those with the weakest unassisted baseline. In the customer-support data, less experienced and lower-skilled agents gained most in speed and quality, while the most skilled saw small gains in speed and small declines in quality (Brynjolfsson, Li and Raymond, 2025). In the dermatology study, the least experienced clinicians gained most from good AI support (Tschandl et al., 2020). In the controlled experiment on ability and calibration, the largest benefit went to people with low ability who knew it (Caplin et al., 2026), and the radiologist replication agreed (Martin, 2026). In the Norwegian trial, CAde raised detection for inexperienced endoscopists and not for experienced ones (Pedersen et al., 2026). The mammography reanalysis found the same structure at the level of individual readers: help for the least discriminating, harm for the most discriminating on hard cases (Povyakalo et al., 2013).

The same clinicians are the ones for whom Q4 and Q5 are most worrying. Inexperienced readers followed incorrect suggestions most readily in the mammography experiment (Dratsch et al., 2023), and learners are the population in which the negative Q3 and Q5 results outside medicine were obtained (Liu et al., 2026; Bastani et al., 2025; Shen and Tamkin, 2026). An aid that is most valuable to the less experienced is therefore also the aid most likely to shape how their unassisted skill develops and how often its errors pass through them. The radiology evidence complicates any simple version of this: experience did not reliably predict who benefited among 140 radiologists, and lower performers did not consistently gain more (Yu et al., 2024). What survives the complication is narrower. The distribution of Q1 across clinicians is uneven and poorly predicted by seniority, and nothing in a trial's average assisted effect says how Q2 to Q5 are distributed across the same clinicians. A deployment that reports only the average assisted gain cannot tell whether its benefit is concentrated among the clinicians whose unassisted development it is also changing.

8.2 What the formation evidence licenses

A synthesis of the skill-formation evidence from outside medicine reached a conclusion close to this paper's: deployment data observe the use of existing skill more readily than the formation of new skill, and current measurement cannot determine whether AI erodes formation at population scale (Fan, 2026). This paper's differences are its focus on experts in clinical deployment, its separation of the five quantities, and its extension in the next section of a control-arm argument

that Budzyn et al. (2025) made for colonoscopy.

9. The control arm has already been exposed

9.1 The argument

Consider a trial that randomises colonoscopies, or days, or rooms, between CAde and no CAde, with the same endoscopists performing both arms. Write U0 for an endoscopist's unassisted performance before any exposure, U1 for their unassisted performance while CAde is in routine use on other cases, and A for their assisted performance. The trial compares A with U1, because its control procedures are performed by endoscopists who are also doing CAde procedures. The quantity a purchaser needs, Q1, compares A with U0. The two differ by U0 minus U1, which is Q2 with its sign reversed.

If Q2 is negative, as Budzyn et al. (2025) report, the trial's control arm is depressed and the trial overstates the assisted gain relative to practice without the aid. If Q2 is positive, as the outage analysis suggests for conversational aids (Brynjolfsson, Li and Raymond, 2025), the control arm is lifted and the trial understates it. Only if Q2 is zero does a within-clinician trial estimate Q1 without bias. The deskillling literature and the efficacy literature are therefore not independent evidence to be weighed against each other: if the first is right about the sign and size of Q2, it describes a bias in the second.

The authors of the Polish study made this argument for colonoscopy and supplied the numbers (Budzyn et al., 2025). Its post-period non-AI colonoscopies were done at centres where colonoscopies were assigned to AI or no AI by examination date, by endoscopists doing both, which is what a within-clinician comparison would use as its control arm. In that period, ADR was 25.3% on AI-assisted procedures against 22.4% on non-AI procedures, an apparent gain of 2.9 points, while the assisted figure was 3.1 points below the pre-AI non-AI baseline of 28.4%. In the authors' words, "the increased ADR shown in many randomized trials may, at least in part, be due to a reduction in ADR among non-AI procedures." This paper's contribution is to write the point as the difference between Q1 and Q2, to extend it beyond colonoscopy, and to derive the prediction of Section 9.2. It does not claim that such a bias occurred in any published trial; it claims that a within-clinician design cannot rule it out, and that the size of the possible bias is of the same order as the gain the AGA panel accepted (6.0 points against 8), with the caveat that the Polish study's diagnostic population is not the population of the screening trials.

9.2 What the argument predicts

The argument makes a prediction that can be checked against existing data. Trials that compare exposed clinicians with unexposed ones, by randomising facilities or clinicians, estimate something close to Q1 and are not exposed to Q2 contamination, provided the control clusters stay unexposed. Trials that randomise cases

within clinicians are exposed to it. If Q2 is negative for CADe, within-clinician trials should report larger gains than cluster trials, other things equal. The Veterans Affairs cluster trial reported a 4.2-point rise against a 0.7-point fall in controls (Dominitz et al., 2026), below the 8-point figure the AGA panel accepted from its systematic review of randomised trials (Sultan et al., 2025), a body of evidence that includes patient-randomised designs such as Repici et al. (2020). That gap is consistent with the prediction. It is also consistent with every other explanation of the efficacy-effectiveness gap, including the absence of a Hawthorne effect outside trials and differences in mucosal exposure and case mix (Putera, Koh and Li, 2026), and the Veterans Affairs study measured availability rather than use. The comparison is suggestive, not a test. A test would stratify the CADe meta-analysis by unit of randomisation, which no published synthesis located for this review has done.

9.3 Scope of the argument

Three limits apply. First, the argument requires that exposure in one arm affects performance in the other within the trial's duration. A trial lasting a few weeks may be too short for any spillover, and the Polish study's three-month window is the shortest exposure period over which a clinical Q2 effect has been reported. Second, the argument concerns the control arm's level, not its trend: a trial whose control arm drifts during the study is biased in proportion to the drift. Third, the argument applies equally to MASAI, where examinations were randomised and radiologists were not masked to group (Lang et al., 2023), if the same radiologists read both arms, and to any reader study in which the same readers see aided and unaided cases in alternation. Reader studies that alternate conditions in close succession are especially exposed, because effects of the short-horizon kind that Liu et al. (2026) observed after ten minutes of assisted work would act within a single session.

9.4 Whether this has been said before

The concern that the comparison arm of a clinical trial may be contaminated when the same clinicians deliver both arms is standard in trial methodology, and it is one of the usual motivations for cluster randomisation. For CADe it has been applied to the deskillling question by Budzyn et al. (2025), who write that the non-AI colonoscopy in randomised trials "could be different from everyday colonoscopy, because ADR of the non-AI assisted colonoscopy could be negatively affected by continuous AI exposure". The correspondence on the Polish study could not be read in full and may develop the point further. This paper claims no priority for the argument; what it adds is the Q1 and Q2 algebra, the extension to MASAI and to reader studies, and the cluster-versus-within-clinician prediction of Section 9.2.

10. The numbers in one place

Table 1 collects the measured values on which the argument rests, one row per result, with the quantity each bears on. Figure 1 plots two sets of them.

Three readings of Table 1 matter for what follows. The colonoscopy Q1 rows (1 to 4) are the best estimated and the most variable, and the variation tracks design: trial pooling gives the largest gain, the cluster trial a smaller one, and the non-randomised pooling an interval that includes zero. The Q2 rows (8 to 10) disagree in sign, and the one negative estimate is the one without a concurrent comparison. The Q4 rows (16 to 18) agree in direction and are large: in the mammography experiment, wrong suggestions cut the share of correct ratings by roughly 45 to 75 percent of its value on mammograms with a correct suggestion, depending on experience (a proxy for Q4, since the two sets of mammograms differ), and in the vignette study a biased model cost 11.3 points of accuracy. On present evidence, the deskillling quantity with the most secure magnitude is the one least often called deskillling.

11. A decision structure

Algorithm 1 states the decision structure the evidence supports. Its first part classifies a study by what its comparison arm saw, which fixes the quantity it estimates. Its second part combines the quantities into the expected performance of a deployed aid, and identifies which terms a given body of evidence leaves unmeasured. It is a synthesis of the results above.

Three features of the procedure deserve comment. Line 8 encodes Section 9: the effect estimated by a within-clinician trial is the assisted effect relative to exposed clinicians. Line 10 makes explicit that unassisted performance enters the value of a deployed aid only through u , the share of unaided cases, and that the aid's own error rate enters through A_{wrong} , where the Q4 evidence lives. If u is small, even a large negative Q2 changes V little; if e is large, a modest increase in automation bias with exposure changes it a great deal. Line 14 is the honest summary of the present literature. Of the terms in line 10, none is estimated separately with any precision for CADe: trials estimate the average of A_{right} and A_{wrong} at the trial's error rate, and $U0$ only from control arms of designs whose comparison clinicians are unexposed; within-clinician control arms estimate $U1$. A_{wrong} has been estimated in reader experiments but not in deployment. $U1$ has three estimates of conflicting sign. And u has not been reported at all.

12. Counter-evidence and the strongest case against this paper

12.1 "Budzyn's result is simply a confounded before-after comparison, so there is nothing to decompose"

On this view the Polish study is an artefact, Pedersen and Okumura are correct, and the decomposition dresses up a single weak result. The objection has force against the Polish study, whose six-month before-after window lacks a concurrent comparison. It does not remove the need for the decomposition. Q4 is large and well replicated whatever happens to Q2, and Q5 has negative

Table 1. Published values bearing on the five quantities. Every value is reported in the source named on the row. Q1 assisted effect; Q2 interleaved withdrawal; Q3 sustained withdrawal; Q4 failure mode; Q5 formation.

#	Result	Value	Setting	Q	Source
1	Pooled ADR, CAde vs control	36.6% vs 25.2% (RR 1.44)	5 RCTs, 4,354 patients	Q1	Hassan et al., 2021
2	ADR gain accepted by panel	+8 points (6 to 10)	guideline synthesis	Q1	Sultan et al., 2025
3	ADR, non-randomised studies	44% vs 38% (RR 1.11, 0.97 to 1.28)	8 studies, 9,782 patients	Q1	Patel et al., 2024
4	ADR change, cluster trial	+4.2 (CAde sites) vs -0.7 (controls)	139 VA facilities	Q1	Dominitz et al., 2026
5	Interval cancers per 1,000	1.55 vs 1.76 (ratio 0.88)	MASAI, ~105,900 women	Q1	Gommers et al., 2026
6	Screen readings	61,248 vs 109,692 (-44.2%)	MASAI	Q1	Hernstrom et al., 2025
7	Within-reader sensitivity, CAD	OR 0.53 (0.29 to 0.97)	107 radiologists	Q1/ Q4	Lehman et al., 2015
8	Non-AI ADR before vs after AI; AI-assisted ADR after	28.4% vs 22.4% (-6.0); assisted 25.3%	4 centres, 1,443 non-AI + 734 AI	Q2	Budzyn et al., 2025
9	Non-CAde ADR over 3 years	maintained, no decline	1 centre, 4,712 matched	Q2	Okumura et al., 2026
10	Chat speed during outages	faster than pre-AI baseline	5,172 support agents	Q2	Brynjolfsson et al., 2025
11	PDR-5, inexperienced	31.9 / 39.5 / 36.3%	before / with / after	Q1/ Q3	Pedersen et al., 2026
12	PDR-5, experienced	28.3 / 26.7 / 27.2%	before / with / after	Q1/ Q3	Pedersen et al., 2026
13	Skill loss after >1 year non-use	d about -1.4	189 data points	Q3	Arthur et al., 1998
14	Unassisted solve rate after AI	0.71 vs 0.77 (d -0.19)	fractions, Exp. 2	Q3/ Q5	Liu et al., 2026
15	Grades after access removed	-17% (GPT Base)	~1,000 students	Q3/ Q5	Bastani et al., 2025
16	Correct ratings, AI wrong	19.8 / 24.8 / 45.5% (vs ~80% when AI right)	27 radiologists	Q4	Dratsch et al., 2023
17	Accuracy, biased model	-11.3 points vs 73.0% baseline	457 clinicians	Q4	Jabbour et al., 2023
18	CAD effect by reader	+0.016 weak; -0.145 strongest	50 readers, reanalysis	Q4	Povyakalo et al., 2013
19	Quiz score after AI learning	-17% (d 0.738)	new Python library	Q5	Shen and Tamkin, 2026

Algorithm 1. Reading a deskillung claim against the deployment value it bears on.

```

Input: study S; deployment with share u of cases done without the aid,
aid error rate e on aided cases, horizon T since introduction
1: classify S by what its comparison arm saw:
2: unexposed clinicians, aid absent          -> estimates Q1
3: unaided cases interleaved with aided ones -> estimates Q2
4: unaided cases after the aid is withdrawn  -> estimates Q3
5: aided cases where the aid is wrong        -> estimates Q4
6: learners trained with vs without the aid  -> estimates Q5
7: if S randomises cases, days or rooms within clinician:
8: mark its Q1 estimate as A - U1, biased by -Q2 relative to A - U0
9: check the threats for its class: concurrent control, secular trend,
case mix, novelty window, power, outcome proxy (ADR vs PDR-5)
10: expected performance with the aid at horizon T:
    V = (1-u) * [ (1-e) * A_right + e * A_wrong ] + u * U1(T)
11: compare V with U0, the unexposed baseline; report which of
    u, e, A_wrong and U1(T) were measured rather than assumed
12: if the aid may be withdrawn: add the Q3 term U_after(T) - U0
13: if trainees use the aid: add the Q5 term for the next cohort
14: if any term is unmeasured: state the decision as conditional on it
Output: the quantity S estimates, its threats, and the open deployment terms

```

experimental evidence outside medicine. The control-arm argument does not depend on the sign of Q2, only on the possibility that it is not zero, and the outage study shows a non-zero Q2 of the opposite sign in another domain. If the Polish result were withdrawn tomorrow, the argument of Sections 3, 7, 9 and 11 would stand.

12.2 "The control-arm bias is too small to matter"

If Q2 is a few tenths of a percentage point, the bias in a within-clinician trial is negligible. That may be so. The only published point estimate for CAde is -6.0 points, with an interval reaching -1.6 at its smaller end, and the only concurrent-comparison estimate is null. The paper's claim is not that the bias is large; it is that the size is unknown, that within-clinician designs cannot estimate

it, and that a meta-analysis stratified by unit of randomisation could bound it with data that already exist.

12.3 "Deskillung is an empirical question about colonoscopy; the human-factors analogies do not transfer"

The aviation and laboratory evidence is used here for structure, not magnitude. It shows that Q3 can be component-specific (Casner et al., 2014), that recent practice of the specific component predicts retention (Haslbeck and Hoermann, 2016; Arthur et al., 1998), and that automation bias survives expertise (Parasuraman and Manzey, 2010; Dratsch et al., 2023). None of these is a claim about the size of any effect in colonoscopy, and the paper does not use them as one.

12.4 "Experienced clinicians are not affected, so the issue is confined to trainees"

The Norwegian trial found no effect of CADe on experienced endoscopists' PDR-5 during use or after removal (Pedersen et al., 2026), and the Japanese study found high detectors improving on non-CADe procedures (Okumura et al., 2026). Against this, the only negative clinical Q2 estimate was obtained in 19 endoscopists with more than 2,000 colonoscopies each (Budzyn et al., 2025), a point Pedersen et al. (2026) themselves make. The mammography experiment found very experienced readers' accuracy falling to 45.5% when the aid was wrong (Dratsch et al., 2023), and the radiology heterogeneity results found experience a poor predictor of who benefits (Yu et al., 2024). Experience appears to protect against some quantities more than others: perhaps against Q3, less against Q4. That is itself an argument for keeping them apart.

12.5 "This is an argument for better trials, which everyone already wants"

Partly. The paper's specific additions are the five-way decomposition, which builds on the distinction between interleaved and sustained withdrawal that Pedersen et al. (2026) draw; the observation that two Q2 studies conflict; a generalisation of the control-arm argument that Budzyn et al. (2025) made for colonoscopy, through the Q1 and Q2 algebra, its extension to MASAI and reader studies, and the checkable cluster-versus-within-clinician prediction; and the identification of u as the missing deployment term. A reader who considers these obvious in retrospect is not being unreasonable. They are not, however, stated together in the reviews located for this paper, and the guideline abstracts summarised in Section 4.2 do not mention them.

13. What would settle the open parts

Three designs, each aimed at one of the open terms, would settle most of what is open. None requires new technology.

A cluster trial with scheduled unassisted audits. Randomise facilities or clinicians, not cases, to receive the aid, and in the exposed arm schedule unassisted

sessions at fixed intervals: a non-AI day every few weeks, recorded as such. The comparison of exposed clinicians' unassisted sessions with unexposed clinicians' routine work estimates Q2 against a concurrent control, free of secular trend. The comparison of the exposed arm's assisted work with the unexposed arm estimates Q1 free of Q2 contamination. Dominitz et al. (2026) has the cluster structure; it lacks the audits.

Randomised withdrawal. Among clinicians who have used the aid for a fixed period, randomise some to have it withdrawn and others to keep it, and follow both on unassisted sessions. The contrast estimates Q3 against a concurrent control, removing the time effect that a fixed-order design like the Norwegian trial cannot separate (Pedersen et al., 2026). Outcomes should be component-specific where possible: withdrawal time and inspection quality alongside detection, in the way the aviation work separated manual control from monitoring (Casner et al., 2014).

Error-injection at two exposure levels. Present the same clinicians with a small number of cases where the aid is deliberately wrong, once at introduction and again after months of routine use, as an audit of automation bias. The change estimates how Q4 moves with exposure, which is the most direct test of whether deskillung matters for safety. Reader experiments already show the method works (Dratsch et al., 2023; Jabbour et al., 2023); what is missing is the repeated measurement.

Alongside these, deployments should log the share of cases handled without the aid, with the reason. This costs almost nothing, and without it no deskillung estimate can be converted into a deployment cost. A meta-analysis of existing CADe trials stratified by unit of randomisation would test the prediction of Section 9.2 with data already published.

14. What is not known

Whether exposure to CADe changes endoscopists' unassisted detection during routine use is not known; three estimates of that quantity disagree in sign, and the one negative estimate lacks a concurrent comparison. Whether it changes what endoscopists can do after the aid is withdrawn is not known for experienced endoscopists, for whom the aid had no effect on PDR-5 during use in the one trial that measured both, although it lengthened their withdrawal times, and is unresolved for less experienced endoscopists because of power and time effects. Whether automation bias grows with exposure is not known, because no study has measured it twice in the same clinicians. Whether trainees who learn colonoscopy with CADe acquire less unassisted skill is not known; there is no formation study in endoscopy. How large the control-arm bias in within-clinician CADe trials is, is not known. The share of cases clinicians handle without an aid once it is deployed is not known for any aid reviewed here. The effect of AI triage on radiologists' long-run reading skill is not known, and the evidence on whether reading volume affects accuracy is itself mixed.

15. Limitations

The most important limitation is access. The Polish study was read in its accepted manuscript, but the five pieces of correspondence around it were not readable in full, so this paper characterises the correspondence only by its existence. The letters may have raised the population question or other points this paper treats as its own inferences, and the authors' reply may answer some of the objections to the before-after design. The MASAI full texts could not be retrieved either, which leaves open whether the same radiologists read both arms.

Second, the evidence base is narrow in the clinical domain that carries most of the argument. There are three colonoscopy studies of unassisted performance after exposure, the one read in full involving 13 endoscopists, and a single mammography trial whose design makes it relevant to Section 9. Several of the most informative sources are 2026 publications or preprints (Pedersen et al., 2026; Okumura et al., 2026; Dominitz et al., 2026; Liu et al., 2026; Shen and Tamkin, 2026; Wu et al., 2026), and the non-clinical experiments are short-horizon and mostly about learners.

Third, the interpretive moves are this paper's. The five-way decomposition is a stipulation. The suggestion that aids supplying content differ from aids supplying pointers is a conjecture with no direct test. The control-arm argument was stated for colonoscopy by Budzyn et al. (2025); its extension to other designs is an inference from those designs, and the comparison of the cluster trial with the meta-analysis in Section 9.2 is consistent with it without testing it.

Fourth, the paper treats detection rates as outcomes. They are surrogates. Both guideline panels concluded that the link from detection gains to fewer cancers is uncertain (Sultan et al., 2025; Foroutan et al., 2025), and the same uncertainty applies to detection losses.

16. Conclusion

The question "does clinical AI deskilling?" has no single answer because it is not a single question. Separated into five quantities, the evidence looks different from the way it is usually summarised. The assisted effect is well estimated and varies with design. Unassisted performance while the aid is in use has three estimates that disagree. Unassisted performance after withdrawal has one clinical trial, which found no change. Performance when the aid is wrong is well measured, consistently large, and rarely called deskilling. Formation has experimental evidence outside medicine and almost none inside it.

Two structural points follow. Trials that randomise cases within clinicians compare the aid against exposed clinicians, so whatever deskilling or upskilling exists is already inside their effect estimates, with a sign set by the sign of the effect. And no study reports the share of work done without the aid, which is the term that turns any deskilling estimate into a cost. Until both are addressed, the honest summary is that the gain is

measured against a baseline the aid may have moved, and the loss is measured on a share of work nobody has counted.

References

- Agarwal, N., Moehring, A., Rajpurkar, P., & Salz, T. (2023). Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology. NBER Working Paper 31422. doi:10.3386/w31422.
- Abdulnour, R.-E. E., Gin, B., & Boscardin, C. K. (2025). Educational Strategies for Clinical Supervision of Artificial Intelligence Use. *New England Journal of Medicine*, 393(8), 786-797. doi:10.1056/NEJMr2503232.
- Ahmad, O. F. (2025). Endoscopist deskilling: an unintended consequence of AI-assisted colonoscopy? *The Lancet Gastroenterology & Hepatology*, 10(10), 872-873. doi:10.1016/S2468-1253(25)00164-5.
- Arthur, W., Jr., Bennett, W., Jr., Stanush, P. L., & McNelly, T. L. (1998). Factors That Influence Skill Decay and Retention: A Quantitative Review and Analysis. *Human Performance*, 11(1), 57-101. doi:10.1207/s15327043hup1101_3.
- Bainbridge, L. (1983). Ironies of automation. *Automatica*, 19(6), 775-779. doi:10.1016/0005-1098(83)90046-8.
- Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakci, O., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26). doi:10.1073/pnas.2422633122.
- Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at Work. *The Quarterly Journal of Economics*, 140(2), 889-942. doi:10.1093/qje/qjae044. Outage analysis read in the arXiv version, arXiv:2304.11771, doi:10.48550/arXiv.2304.11771.
- Budzyn, K., Romanczyk, M., Kitale, D., Kolodziej, P., Bugajski, M., Adami, H. O., et al. (2025). Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. *The Lancet Gastroenterology & Hepatology*, 10(10), 896-903. doi:10.1016/S2468-1253(25)00133-5.
- Budzyn, K., Romanczyk, M., & Mori, Y. (2025). Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy - Authors' reply. *The Lancet Gastroenterology & Hepatology*, 10(12), 1062. doi:10.1016/S2468-1253(25)00324-3.
- Buist, D. S. M., Anderson, M. L., Haneuse, S. J., Sickles, E. A., Smith, R. A., Carney, P. A., et al. (2011). Influence of Annual Interpretive Volume on Screening Mammography Performance in the United States. *Radiology*, 259(1), 72-84. doi:10.1148/radiol.10101698.
- Caplin, A., Deming, D., Li, S., Martin, D., Marx, P., Weidmann, B., & Ye, K. J. (2026). The ABCs of Who Benefits from Working with AI: Ability, Beliefs, and Calibration. *Management Science*, 72(7), 5843-5852. doi:10.1287/mnsc.2024.08994.
- Casner, S. M., Geven, R. W., Recker, M. P., & Schooler, J. W. (2014). The Retention of Manual Flying Skills in the Automated Cockpit. *Human Factors*, 56(8), 1506-1516. doi:10.1177/0018720814535628.
- Dominitz, J. A., Gawron, A. J., McKee, G. B., Hoggatt, K. J., & Kaltenbach, T. (2026). Impact of Availability of Computer-Aided Detection Devices on Adenoma Detection During Colonoscopy: A Cluster Randomized Study. *Gastroenterology*. doi:10.1053/j.gastro.2026.05.018.
- Dratsch, T., Chen, X., Rezazade Mehrizi, M., Kloeckner, R., Mahringer-Kunz, A., Pusken, M., et al. (2023). Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance. *Radiology*, 307(4). doi:10.1148/radiol.222176.
- Ebbatson, M., Harris, D., Huddleston, J., & Sears, R. (2010). The relationship between manual handling performance and recent flying experience in air transport pilots. *Ergonomics*, 53(2), 268-277. doi:10.1080/00140130903342349.

- Eisemann, N., Bunk, S., Mukama, T., Baltus, H., Elsner, S. A., Gomille, T., et al. (2025). Nationwide real-world implementation of AI for cancer detection in population-based mammography screening. *Nature Medicine*, 31(3), 917-924. doi:10.1038/s41591-024-03408-6.
- Endsley, M. R., & Kiris, E. O. (1995). The Out-of-the-Loop Performance Problem and Level of Control in Automation. *Human Factors*, 37(2), 381-394. doi:10.1518/001872095779064555.
- Fan, A. X. (2026). Toward Measuring AI's Effects on Skill Formation: The Stock-Formation Gap. arXiv:2605.16283. doi:10.48550/arXiv.2605.16283. [PREPRINT]
- Fenton, J. J., Taplin, S. H., Carney, P. A., Abraham, L., Sickles, E. A., D'Orsi, C., et al. (2007). Influence of Computer-Aided Detection on Performance of Screening Mammography. *New England Journal of Medicine*, 356(14), 1399-1409. doi:10.1056/NEJMoa066099.
- Foroutan, F., Vandvik, P. O., Helsingen, L. M., Kalager, M., Rutter, M., Selby, K., et al. (2025). Computer aided detection and diagnosis of polyps in adult patients undergoing colonoscopy: a living clinical practice guideline. *BMJ*, 388, e082656. doi:10.1136/bmj-2024-082656.
- Gaessler, F., & Piezunka, H. (2023). Training with AI: Evidence from chess computers. *Strategic Management Journal*, 44(11), 2724-2750. doi:10.1002/smj.3512.
- Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., et al. (2021). Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digital Medicine*, 4. doi:10.1038/s41746-021-00385-9.
- Gerke, S., Hassan, C., & Mori, Y. (2026). Human deskilling in medical artificial intelligence: prohibited or permissible under the EU Artificial Intelligence Act? *Nature Reviews Gastroenterology & Hepatology*, 23(7), 521-522. doi:10.1038/s41575-026-01210-y.
- Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1), 121-127. doi:10.1136/amiajnl-2011-000089.
- Gommers, J., Hernstrom, V., Josefsson, V., Sartor, H., Schmidt, D., Hjelmgren, A., et al. (2026). Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study: a randomised, controlled, non-inferiority, single-blinded, population-based, screening-accuracy trial. *The Lancet*, 407, 505-514. doi:10.1016/S0140-6736(25)02464-X.
- Hassan, C., Spadaccini, M., Iannone, A., Maselli, R., Jovani, M., Chandrasekar, V. T., et al. (2021). Performance of artificial intelligence in colonoscopy for adenoma and polyp detection: a systematic review and meta-analysis. *Gastrointestinal Endoscopy*, 93(1), 77-85. doi:10.1016/j.gie.2020.06.059.
- Haslbeck, A., & Hoermann, H.-J. (2016). Flying the Needles: Flight Deck Automation Erodes Fine-Motor Flying Skills Among Airline Pilots. *Human Factors*, 58(4), 533-545. doi:10.1177/0018720816640394.
- Hernstrom, V., Josefsson, V., Sartor, H., Schmidt, D., Larsson, A.-M., Hofvind, S., et al. (2025). Screening performance and characteristics of breast cancer detected in the Mammography Screening with Artificial Intelligence trial (MASAI): a randomised, controlled, parallel-group, non-inferiority, single-blinded, screening accuracy study. *The Lancet Digital Health*, 7(3), e175-e183. doi:10.1016/S2589-7500(24)00267-X.
- Heudel, P. E., Crochet, H., Filori, Q., Bachelot, T., & Blay, J. Y. (2026). Artificial intelligence in medicine: a scoping review of the risk of deskilling and loss of expertise among physicians. *ESMO Real World Data and Digital Oncology*, 12, 100693. doi:10.1016/j.esmorw.2026.100693.
- Jabbour, S., Fouhey, D., Shepard, S., Valley, T. S., Kazerooni, E. A., Banovic, N., et al. (2023). Measuring the Impact of AI in the Diagnosis of Hospitalized Patients: A Randomized Clinical Vignette Survey Study. *JAMA*, 330(23), 2275. doi:10.1001/jama.2023.22295.
- Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task. arXiv:2506.08872. doi:10.48550/arXiv.2506.08872. [PREPRINT]
- Ladabaum, U., Shepard, J., Weng, Y., Desai, M., Singer, S. J., & Mannalithara, A. (2023). Computer-aided Detection of Polyps Does Not Improve Colonoscopist Performance in a Pragmatic Implementation Trial. *Gastroenterology*, 164(3), 481-483. doi:10.1053/j.gastro.2022.12.004.
- Lam, K. (2025). Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy. *The Lancet Gastroenterology & Hepatology*, 10(12), 1060. doi:10.1016/S2468-1253(25)00290-0.
- Lang, K., Josefsson, V., Larsson, A.-M., Larsson, S., Hogberg, C., Sartor, H., et al. (2023). Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *The Lancet Oncology*, 24(8), 936-944. doi:10.1016/S1470-2045(23)00298-X.
- Lehman, C. D., Wellman, R. D., Buist, D. S. M., Kerlikowske, K., Tosteson, A. N. A., & Miglioretti, D. L. (2015). Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. *JAMA Internal Medicine*, 175(11), 1828. doi:10.1001/jamainternmed.2015.5231.
- Levartovsky, A., & Kopylov, U. (2025). Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy. *The Lancet Gastroenterology & Hepatology*, 10(12), 1060-1061. doi:10.1016/S2468-1253(25)00289-4.
- Liu, G., Christian, B., Dumbalska, T., Bakker, M. A., & Dubey, R. (2026). AI Assistance Reduces Persistence and Hurts Independent Performance. *COLM 2026*. arXiv:2604.04721. doi:10.48550/arXiv.2604.04721.
- Lyell, D., & Coiera, E. (2017). Automation bias and verification complexity: a systematic review. *Journal of the American Medical Informatics Association*, 24(2), 423-431. doi:10.1093/jamia/ocw105.
- Maas, M. H. J., Rath, T., Spada, C., Soons, E., Forbes, N., Kashin, S., et al. (2024). A computer-aided detection system in the everyday setting of diagnostic, screening, and surveillance colonoscopy: an international, randomized trial. *Endoscopy*, 56, 843-850. doi:10.1055/a-2328-2844.
- Macnamara, B. N., Berber, I., Cavusoglu, M. C., Krupinski, E. A., Nallapareddy, N., Nelson, N. E., Smith, P. J., Wilson-Delfosse, A. L., & Ray, S. (2024). Does using artificial intelligence assistance accelerate skill decay and hinder skill development without performers' awareness? *Cognitive Research: Principles and Implications*, 9. doi:10.1186/s41235-024-00572-8.
- Martin, D. (2026). Revisiting the ABCs of Working with AI: A Replication with Radiologists. arXiv:2606.12585. doi:10.48550/arXiv.2606.12585. [PREPRINT]
- Mori, Y., Kopylov, U., Sinonquel, P., Ebigbo, A., Dekker, E., De Groof, A. J., et al. (2026). Curriculum for safe and effective use of artificial intelligence in endoscopy: European Society of Gastrointestinal Endoscopy (ESGE) Position Statement. *Endoscopy*, 58, 202-210. doi:10.1055/a-2742-4342.
- Natali, C., Marconi, L., Dias Duran, L. D., & Cabitza, F. (2025). AI-induced Deskilling in Medicine: A Mixed-Method Review and Research Agenda for Healthcare and Beyond. *Artificial Intelligence Review*, 58(11). doi:10.1007/s10462-025-11352-1.
- Okumura, T., Kudo, S.-E., Ide, Y., Kato, S., Miyata, Y., Takisima, K., et al. (2026). Long-term impact of computer-aided adenoma detection: a prospective observational study. *Endoscopy*, 58(2), 121-129. doi:10.1055/a-2661-2624.
- Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human Performance Consequences of Stages and Levels of Automation: An Integrated Meta-Analysis. *Human Factors*, 56(3), 476-488.

- doi:10.1177/0018720813501549.
- Parasuraman, R., & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors*, 52(3), 381-410. doi:10.1177/0018720810376055.
- Patel, H. K., Mori, Y., Hassan, C., Rizkala, T., Radadiya, D. K., Nathani, P., et al. (2024). Lack of Effectiveness of Computer Aided Detection for Colorectal Neoplasia: A Systematic Review and Meta-Analysis of Nonrandomized Studies. *Clinical Gastroenterology and Hepatology*, 22, 971-980. doi:10.1016/j.cgh.2023.11.029.
- Pedersen, T. A., Mori, Y., Botteri, E., Engjom, T., Seip, B., Dimcevski, G. G., & Havre, R. F. (2026). Learning and deskillling effects of artificial intelligence in colonoscopy among endoscopists with different levels of experience: a pragmatic, prospective trial. *Endoscopy*, 58(9), 1003-1014. doi:10.1055/a-2858-7084.
- Povyakalo, A. A., Alberdi, E., Strigini, L., & Ayton, P. (2013). How to Discriminate between Computer-Aided and Computer-Hindered Decisions: A Case Study in Mammography. *Medical Decision Making*, 33(1), 98-107. doi:10.1177/0272989X12465490.
- Putera, M., Koh, F. H. X., & Li, J. W. (2026). Controversies in Computer-Assisted Detection in Colonoscopy. *Digestion*. doi:10.1159/000550268.
- Repici, A., Badalamenti, M., Maselli, R., Correale, L., Radaelli, F., Rondonotti, E., et al. (2020). Efficacy of Real-Time Computer-Aided Detection of Colorectal Neoplasia in a Randomized Trial. *Gastroenterology*, 159(2), 512-520. doi:10.1053/j.gastro.2020.04.062.
- Shen, J. H., & Tamkin, A. (2026). How AI Impacts Skill Formation. arXiv:2601.20245. doi:10.48550/arXiv.2601.20245. [PREPRINT]
- Shin, M., Kim, J., van Opheusden, B., & Griffiths, T. L. (2023). Superhuman artificial intelligence can improve human decision-making by increasing novelty. *Proceedings of the National Academy of Sciences*, 120(12). doi:10.1073/pnas.2214840120.
- Sultan, S., Shung, D. L., Kolb, J. M., Foroutan, F., Hassan, C., Kahi, C. J., et al. (2025). AGA Living Clinical Practice Guideline on Computer-Aided Detection-Assisted Colonoscopy. *Gastroenterology*, 168(4), 691-700. doi:10.1053/j.gastro.2025.01.002.
- Topaz, M., Peltonen, L. M., Zhang, Z., & Backonja, U. (2026). The Deskillling Effect: Is Artificial Intelligence Eroding Clinical Competence? *Annals of Internal Medicine*, 179(7), 1048-1050. doi:10.7326/ANNALS-26-00613.
- Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., et al. (2020). Human-computer collaboration for skin cancer recognition. *Nature Medicine*, 26, 1229-1234. doi:10.1038/s41591-020-0942-0.
- van de Sande, D., Sleijfer, S., Buijsman, S. N. R., Gommers, D., & van Genderen, M. E. (2025). Endoscopist deskillling risk after exposure to artificial intelligence in colonoscopy. *The Lancet Gastroenterology & Hepatology*, 10(12), 1061-1062. doi:10.1016/S2468-1253(25)00288-2.
- Watkins, K., Ladabaum, U., Olsen, E., Hoogerbrug, J., Mannalithara, A., Weng, Y., Shaw, B., Bohn, R., & Singer, S. (2024). Exploring human-artificial intelligence interactions in a negative pragmatic trial of computer-aided polyp detection. *iGIE*, 3(2), 274-285. doi:10.1016/j.igie.2024.04.016.
- Wei, M. T., Shankar, U., Parvin, R., Abbas, S. H., Chaudhary, S., Friedlander, Y., et al. (2023). Evaluation of Computer-Aided Detection During Colonoscopy in the Community (AI-SEE): A Multicenter Randomized Clinical Trial. *American Journal of Gastroenterology*, 118(10), 1841-1847. doi:10.14309/ajg.0000000000002239.
- Wu, S., Belem, C. G., Fu, S., Steyvers, M., & Smyth, P. (2026). How AI Assistance Affects Human Skill Development: A Study of Learning with Logic Puzzles. arXiv:2608.23543. doi:10.48550/arXiv.2608.23543. [PREPRINT]
- Yao, S. (2026). When the Scaffold Stays On: AI, Practice Style, and Screening in Elite Skill Formation. arXiv:2606.06253. doi:10.48550/arXiv.2606.06253. [PREPRINT]
- Yu, F., Moehring, A., Banerjee, O., Salz, T., Agarwal, N., & Rajpurkar, P. (2024). Heterogeneity and predictors of the effects of AI assistance on radiologists. *Nature Medicine*, 30(3), 837-849. doi:10.1038/s41591-024-02850-w.