How AI Interface Design Shapes Independent Judgment After Assistance Is Removed

CHRISTOPH TRATTNER, SFI MediaFutures, University of Bergen, Norway
TOBIAS JOVALL WESSEL, SFI MediaFutures, University of Bergen, Norway
YELYZAVETA LYSOVA, SFI MediaFutures, University of Bergen, Norway
ALAIN STARKE, SFI MediaFutures, University of Bergen, Norway

Artificial intelligence (AI) can help users verify health claims, but may inhibit them to actively engage with the content. It's uncertain whether AI assistance builds skill for future judgments without it. We examined whether requiring people to inspect evidence and commit to a judgment before seeing an AI-generated answer improved later unaided verification. 474 U.S.-based participants judged the veracity of health claims in a two-wave experiment. In Wave 1, they faced either an AI-generated answer before showing source information ('answer-first AI'), the evidence before the answer ('evidence-first AI'), or received no AI assistance ('source practice'). A week later (Wave 2), participants judged 16 new claims and its evidence without AI assistance. We did not observe differences between the different modes of AI assistance, nor when compared to the no-AI baseline. We discuss other effects, based on time spent and immediate claim judgment accuracy, as well as the study's implications.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; *Interactive systems and tools*.

Additional Key Words and Phrases: human-AI interaction, AI-assisted decision making, interface design, fact-checking, misinformation, learning, claim verification, longitudinal study

1 Introduction

Ask a chatbot whether "does zinc lozenges shorten a cold?", and it will provide an answer, an explanation, and a source within seconds. AI assistance can help people to make better decisions when judging the veracity of a claim or statement [1, 24]. Most research in this area focuses on how user skills or literacy levels lead to factchecking behavior (cf. [18, 40]). However, would there be an effect the other way around, where extensively using AI to verify claims support a user's skills or capabilities to examine further claims? Moreover, are users better equipped to verify a health claim independently after using AI-generated factchecks extensively, when AI assistance is unavailable? Or, is the accuracy of their judgment merely 'on loan', disappearing once AI assistance is gone?

These questions touch upon how assistance is provided in the user interface, depending on how and when answers and evidence are shown. We consider this analogous to worked-example effect in education [29], where teachers lower the cognitive load among students by showing how a problem can be solved rather than letting them solve it themselves (cf. [13, 27]. For AI assistants, one pattern would be to show the AI's answer first, while highlighting the evidence later. A different method, which is recommended to keep people "in the loop," would hide the answer until the user has read the evidence, chosen a source, and committed to an accuracy judgment (cf. [4, 6]). Requiring a decision before advice reduces how often people accept wrong AI suggestions during use [6], and decision order changes how advice is used [11]. However, whether

Authors' Contact Information: Christoph Trattner, SFI MediaFutures, University of Bergen, Norway; Tobias Jovall Wessel, SFI MediaFutures, University of Bergen, Norway; Yelyzaveta Lysova, SFI MediaFutures, University of Bergen, Norway; Alain Starke, SFI MediaFutures, University of Bergen, Norway.

this setup also leaves users more capable for future judgments of different claims is tested in this paper, examining whether delaying the disclosure of a correct answer would build verification skills.

In earlier studies, the relation AI assistance and learning is unclear. Gajos and Mamykina [12] show that choosing before an AI recommendation does not produce incidental learning in a nutrition task, whereas an explanation without a recommendation does so. In different domains, Buçinca et al. [4] shows that explanations written to address a user's likely misconception improves unaided decisions, while in education, students who practice with unrestricted AI seem to perform worse than students without it once access to AI is removed [2]. What helped learning in these studies is changing the context in which the AI assistance was provided. For the same evidence, correct answer, and explanation, they differed in whether a user had to commit to a judgment before seeing it.

This paper reports on an experiment where participants judged the veracity of health claims, using different forms of AI assistance. After six baseline claims judged without AI, participants practiced on ten claims with three evidence cards each. *Answer-first AI* showed a correct, scripted answer and explanation before the participant judged. *Evidence-first AI* required a source choice and an initial judgment before showing the same answer, then asked for a final judgment. *Source-practice* required the same evidence work with an answer key and no AI. Every AI answer was correct, being prevalidated by expert judgment. An eight-claim immediate test followed, while participants also returned about a week later to judge sixteen new claims with evidence cards and no AI.

The paper makes three contributions. (1) A randomized two-wave test, with matched evidence and an always-correct assistant, of whether requiring people to decide before AI advice improves later unaided verification: it did not, and the interval rules out the benefit designers hope for. (2) Evidence that felt control and later capability come apart: the interface that made people commit first raised their sense of ownership and their reported deference, and lowered their unaided accuracy. (3) A worked example of evaluating AI assistance by what remains after it, including an implementation deviation, the audit that found it, and the bounded analysis that shows what it does and does not change.

The epistemic-transfer framework motivates separating assisted success from later performance [35]. A revision of that non-archival preprint briefly summarizes selected outcomes from the same cohort. The present paper provides the complete empirical design, analysis, and supplementary outcomes.

2 Related Work

2.1 Use of AI assistance

Research on AI assistance is versatile, but tends to be conducted in individual experiments. Studies have shown that people sometimes reject useful AI advice after seeing errors [10], and sometimes follow incorrect advice [19, 22, 25]. One approach to mitigate this is the use of cognitive forcing functions, which inhibit heuristic overreliance on assistive tools [4]. One method is to setup a context in which a decision is required from a user before advice is shown, which reduced accepted errors at a cost to interface preference [6]. Such decision order also matters in clinical imaging [11], where users who have to commit to an answer before receiving assistance are more likely disagree with AI than those who are assisted first. Explanations can support checking when verification is easier than acceptance [36]; partial explanations [8] and critical-thinking prompts [34] can further change advice use.

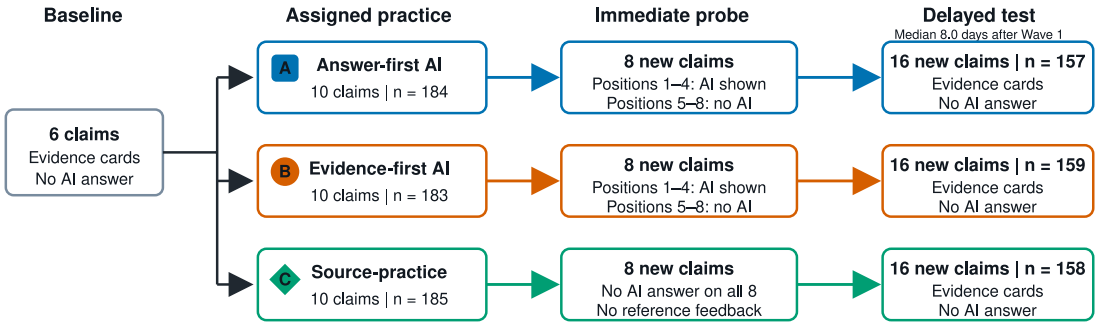


Fig. 1. Study design and sample. All users first judged the veracity of six claims without AI, then practiced on ten claims in one of three randomly assigned interfaces; they remained in these conditions across Wave 1 (Baseline-Immediate probe) and Wave 2 (Delayed test). An immediate test used eight new claims in participant-specific shuffled order; in both AI groups, the answer was always shown on positions 1–4 and absent on positions 5–8. Source-practice had no AI on all eight trials. About a week later, participants judged 16 new claims with evidence cards but no AI answer. Participant retention was 85.9%, which did not differ significantly across conditions.

These studies primarily evaluate decisions while advice is available. We instead evaluate post-assistance performance as the primary outcome. HCI theory distinguishes interface signals from user action [32], and verification actions can change trust without a detectable change in reported agency [20]. Required action, experience, and later performance therefore need separate measures.

2.2 Performance with AI does not predict learning

A growing part of literature now measures performance after AI support is withdrawn. Gajos and Mamykina [12] found that recommendations with explanations improved food choices in the moment but did not teach the underlying rule, whereas reasoning from an explanation *without* a recommendation did. Karny et al. [14] varied when assistance was available in a dynamic decision task and found immediate gains without detectable skill degradation or carryover . Other studies report costs and associations in other learning tasks. Software developers learning a new programming library with AI support showed weaker conceptual understanding, code reading, and debugging ability when the AI tool was removed [30]. In another study, brief AI assistance improved short-term performance but reduced later unassisted performance and persistence across reasoning and comprehension tasks [21]. For logic puzzles, greater reliance on on-demand AI assistance was associated with weaker post-removal skill gains [39]. Similar patterns were found in schools, where unrestricted AI use improved practice performance but harmed later unaided performance unless the tutor withheld answers and provided hints [2]. An observational secondary-school panel study likewise reported a gap between AI-supported homework and closed-book exams [31]. Buçinca et al. further showed that explanations targeting a user's likely misconception can improve unaided decisions [4]. Together, this work establishes post-assistance performance as a distinct outcome. Our narrower question is whether an evidence-first sequence changes it when the evidence, answer, and explanation content are held fixed.

Two lines of learning research suggest competing reasons why one sequence might be useful. Research on retrieval practice says that trying to answer before being told helps, even when the attempt fails [15], and that confidently held errors are corrected especially well by feedback [5]. That is the intuition behind

evidence-first designs. Worked-example research provides a different rationale for beginners: studying a worked solution can reduce the demands of unguided problem solving and help learners identify a solution pattern [33]. An answer-first display resembles a worked example. These accounts motivate the sequence comparison; the experiment does not separately manipulate or identify either learning mechanism. Source-practice provides an active-practice comparison: participants did the evidence work on their own and then received the same reference answer without an AI framing. Because every group practiced, it is not a no-practice control and cannot identify absolute learning relative to receiving no practice.

Buçinca et al. found that explanations contrasting the AI's answer with a predicted plausible alternative improved immediate independent performance relative to explanations justifying only the AI's answer [4]. Their variants using a predicted alternative before a decision or the participant's own choice after a decision did not differ detectably on the reported objective outcomes. That comparison changed the alternative as well as when it was presented, and concerned short-term learning. In misinformation verification, Rani et al. followed 67 participants over four weeks of AI dialogues and found that high assisted accuracy did not translate into sustained accuracy on new items [28]. Their design lacked a longitudinal no-AI comparison, so the observed change does not isolate AI-caused skill loss. Our study instead estimates the delayed difference between randomized, matched-content interface sequences. It does not introduce delayed verification as a new outcome or estimate absolute learning.

2.3 What claim-checking asks of people

Judging an online health claim is not recall. It means noticing where a statement comes from, comparing sources, telling evidence from advertising, and deciding how far a source actually supports the conclusion [23, 26, 38]. An AI assistant shortens that work: it finds sources and collapses them into an answer. That help is not automatically good. DeVerna et al. found that fact-checks from a large language model that correctly spotted most false headlines did not make people better judges and sometimes made them worse [9]. The model can turn a multi-step task into a single decision: accept or reject its answer.

2.4 Contribution

Prior work already distinguishes assisted success from learning, and already tests whether people should decide before advice. The remaining question here is specific: how do these two matched-content sequences compare when people return to judge unfamiliar health claims with evidence but no AI answer? Table 1 locates this comparison relative to the closest studies. The delayed test separates the outcome from the assisted session and uses different claims, while the randomized interface assignment supplies a comparison of the complete experiences. It does not by itself identify absolute learning or a psychological mechanism.

3 Main question and hypotheses

The primary research question examines to what extent the order in which AI judgment and evidence are shown to a user affects its claim judgment accuracy over time. We primarily test the following hypotheses:

[H1]: On new claims judged without the AI answer one week after practice, evidence-first participants are more accurate in their claim assessments than answer-first participants, also when adjusting for baseline accuracy.

Table 1. Closest prior studies and the evidence added here. The studies differ in tasks, interventions and outcomes; their effects are not directly interchangeable.

Study	Question and assessment	Relation to this study
Gajos and Mamykina [12]	Initial choice before advice, with same-session incidental learning in a nutrition task.	Establishes the preceding-decision question; our endpoint is unfamiliar health claims after about a week.
Buçinca et al. [4]	Explanation content and presentation variants, with immediate independent exercise decisions.	Shows why explanation content matters; our reference content is matched across full interface sequences.
Rani et al. [28]	Four weeks of AI dialogue and news discernment, without a longitudinal no-AI comparison.	Establishes delayed verification as an existing concern; our randomized contrast compares two interface sequences.
Present study	Delayed novel-claim verification after answer-first, evidence-first, or active source-practice.	Estimates the complete-sequence difference and its uncertainty relative to the specified practical margin.

In addition, we expect similar advantageous effects immediately after practicing with the sources underlying the AI assistance. We posit the following:

[H2]: Evidence-first participants are in the short-term more accurate in their claim assessments than answer-first participants, also when adjusting for baseline accuracy.

Furthermore, we examined to what to which participants across the two AI conditions changed in their performance when judging claims between the two waves of data collection. We formulate the following hypothesis about this tool removal cost:

[H3]: The difference between final accuracy when the assistant is available and accuracy when it is unavailable will be larger in answer-first than in evidence-first.

4 Method

4.1 Study design

We conducted a between-subjects online experiment with three interface conditions and two waves. In Wave 1, participants completed a short questionnaire, judged six baseline claims without AI assistance, completed ten practice claims using their assigned interface, and judged eight new claims in an immediate test. About one week later, participants completed Wave 2. They judged 16 new claims using evidence cards without an AI answer. Figure 1 summarizes the study sequence. Participants were assigned to one of three conditions for the ten practice claims. In the *answer-first AI* condition, the AI answer appeared before participants reviewed the evidence and made their judgment. In the *evidence-first AI* condition, participants first reviewed the evidence and recorded an initial judgment. They then saw the AI answer and could provide a final judgment. In the *source-practice* condition, participants worked with the evidence without seeing an AI answer. The claims, evidence cards, source-selection task, and reference answers were identical across conditions.

4.2 Participants

The final sample included 474 participants who completed both waves. This corresponds to an 85.9% retention rate from the Wave 1 sample of 552 participants. Participants were recruited through crowdsourcing platform Prolific on 24 and 25 August 2026 for Wave 1. Participants were required to be U.S.-born citizens and current residents. We also required participants to have a prolific approval rate of at least 95% and to have completed a minimum of 100 previous prolific submissions. Their ages ranged from 19 to 80 years ($M = 41.9$, $SD = 12.7$). The sample included 246 women (51.9%), 223 men (47.0%), four non-binary participants (0.8%), and one participant who preferred to self-describe (0.2%). The largest education group held a bachelor's degree ($n = 199$, 42.0%). Median completion time was 19.5 minutes in Wave 1 and 9.8 minutes in Wave 2. Participants received £5.00 for Wave 1 and £2.40 for Wave 2. Among the participants, 234 (49.4%) used AI daily or almost daily, 373 (78.7%) used it at least weekly, and 385 (81.2%) had used AI to check whether information was true.

4.3 Claims and evidence

We created 40 short English health claims: six for the baseline task, ten for practice, eight for the immediate test, and 16 for the delayed test. Topics included antibiotics, sleep, physical activity, vitamins and supplements, sunscreen, detox products, colloidal silver, probiotics, and diet and cancer. Each claim was labeled TRUE, FALSE, or MISLEADING. We defined a misleading claim as one that contained some truth but encouraged an incorrect conclusion. Each claim was linked to an authoritative public source, such as the CDC, WHO, FDA, NIH Office of Dietary Supplements, NIH NCCIH, or National Cancer Institute, and existing health fact-checking datasets [16, 37].

Two members of the team independently classified all 40 claims. They did not see the intended labels or phase assignments. They also rated their confidence, checked the evidence summaries against the sources, and flagged unclear wording or implausible distractors. They agreed on 33 claims (82.5%; Cohen's $\kappa = .70$ [7]). All seven disagreements concerned the distinction between FALSE and MISLEADING. Those were adjudicated, reviewed against the original sources and received wording or evidence-card edits.

4.4 Procedure

After providing consent, participants reported their age, gender, education, frequency of AI use, previous use of AI for fact-checking, general trust in AI, and confidence in identifying false health claims. They also answered five questions about their information-checking habits. Participants first judged six baseline claims. Each screen presented a claim and three evidence cards 2. Participants selected the best source, classified the claim as true, false, or misleading, and reported confidence on a 0-100 slider. The slider ranged from "guessing" to "certain". Participants did not receive an AI answer or corrective feedback. Claim order was randomized, and response times were recorded. Participants were then assigned to a condition and completed ten practice claims using the sequence described on figure 1. All conditions received the same correct reference content, but its timing and framing differed. A short attention check followed the practice task ("select False and enter 80").

The immediate test contained eight new claims in participant-specific random order. In both AI conditions, the assistant's answer appeared on positions 1-4 and was absent on positions 5-8. Answer-first

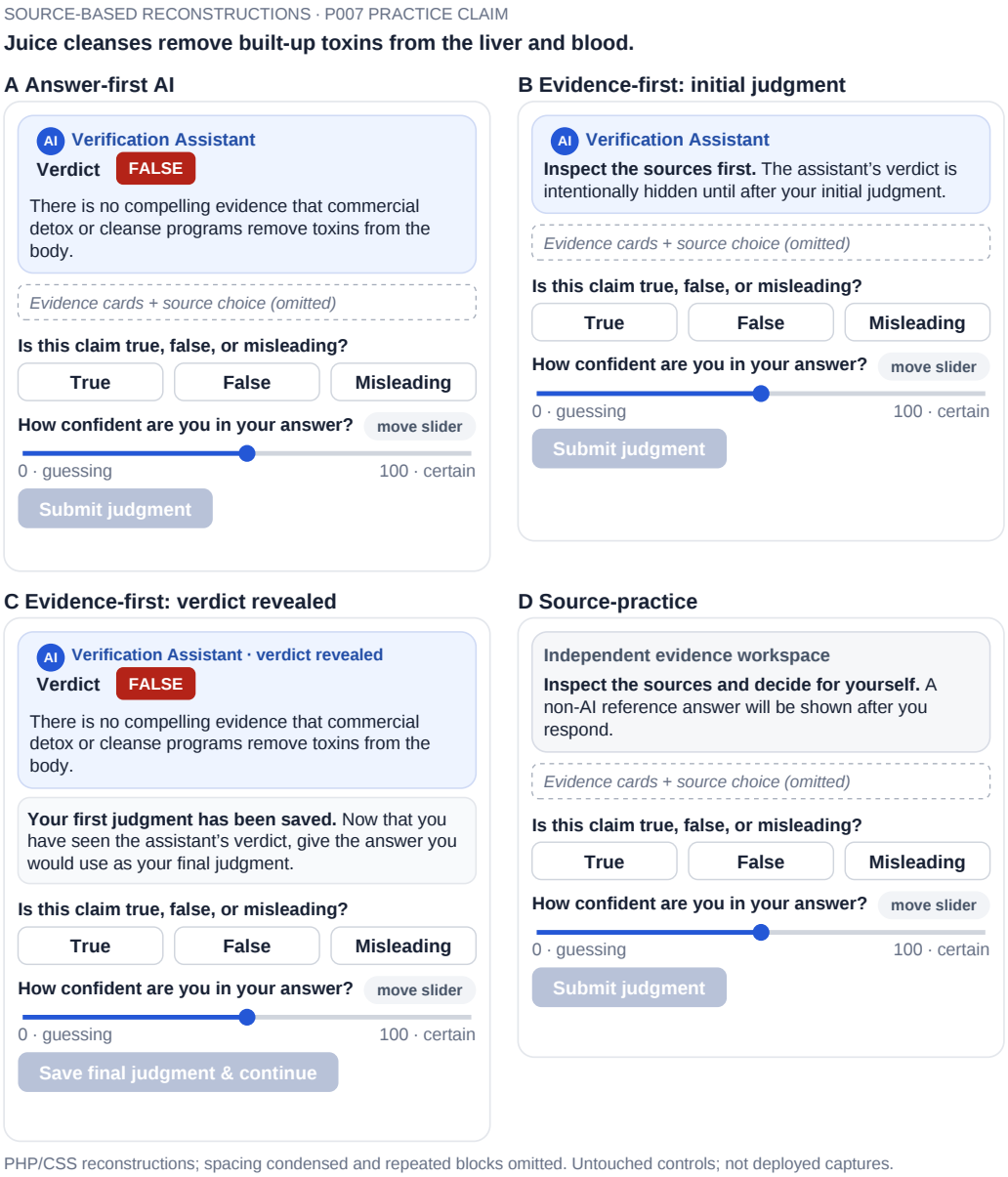


Fig. 2. Source-based reconstructions for practice claim P007: (A) answer-first; (B) evidence-first before commitment; (C) evidence-first after answer revelation; and (D) source-practice. Wording, component order, and initial control states follow the supplied PHP/CSS. Shared claim boxes, evidence-card content, and page chrome are omitted, and spacing is condensed. These are reconstructed illustrations, not deployed-session screenshots.

participants saw the answer before judging the first four claims. Evidence-first participants made an initial judgment, saw the answer, and could revise their response. Both AI groups judged positions 5-8 without

an AI answer. Source-practice participants judged all eight claims without an AI answer. Evidence cards remained available on every trial. No answer-key feedback was provided during the immediate test.

At the end of Wave 1, participants completed 1–7 experience ratings and reported any use of outside tools. Participants were reminded that the claims were not medical advice and that some were false or misleading. We delayed the complete correction sheet until after the follow-up to avoid providing additional instruction before the delayed test.

In Wave 2, participants judged 16 new claims using evidence cards without AI support. Participants were instructed not to use search engines, AI tools, or other outside help during the study.

4.5 Measures

The primary measure was correctness of each judgment, particularly on the delayed test (i.e., Wave 2). Its comparison adjusts for baseline accuracy and claim characteristics. The secondary immediate outcome uses judgments without an AI answer: positions 5–8 in the AI groups, while the source-practice benchmark pools all eight. Assisted accuracy uses final judgments after any required revision. The descriptive block gap is accuracy on early assisted positions minus later unaided positions; it is not an identified causal removal cost.

Each judgment also used a 0–100 confidence slider anchored "guessing" and "certain." This is a rating, not an explicitly elicited probability. The supplementary analyses retain the protocol's squared rating–error proxy, source choices, response times, revisions, and study-specific experience items, with their measurement limitations. They do not establish calibration, validated agency, or a mechanism for the delayed contrast.

4.6 Exclusions and data quality

We only excluded participants with no consent, duplicate or fraudulent records, or no usable task data. These rules were applied before the outcome analyses and without reference to condition-level results; nobody among the 552 Wave 1 completers met them, and all 552 entered the analysis. We did not exclude anyone for failing the instruction check or for self-reported outside help, because both came after assignment. A prespecified sensitivity analysis drops those participants plus anyone whose median response time was below 1.5 seconds. In the final data, one participant failed the instruction check, but was retained as it was accidental. Moreover, eight participants reported outside help in at least one wave (four in each), and nobody was faster than the speed threshold (the fastest median was 4.4 seconds per claim).

4.7 Analysis

The study protocol specified the confirmatory analyses before data collection. Some executable analysis files were incomplete in the archived project materials. They were subsequently reconstructed from the documented model specifications and validated against the archived model outputs. Because the study was not externally preregistered, the timing of these specifications is supported by dated internal project records rather than an independent registry.

Primary comparison. We fitted the prespecified mixed-effects logistic model using `lme4` [3]. The unit of analysis was a judgment, with random intercepts for participant and claim:

$$\operatorname{logit} P(\text{correct}_{ij}) = \beta_0 + \beta_1 \text{Interface}_i + \beta_2 \text{Baseline}_i + \beta_3 \text{Label}_j + \beta_4 \text{Order}_{ij} + u_i + v_j.$$

The primary contrast was evidence-first versus answer-first, h odds ratios below 1 favor answer-first. H1 uses the 474 observed delayed returners. Random assignment supports comparison of the complete sequences, but missing outcomes still require assumptions: balanced return rates cannot exclude selective attrition. The sensitivity checks below address recorded exclusions and timing, not arbitrary missing-not-at-random mechanisms.

Three distinct accuracy summaries. Raw accuracy describes the observed participants and sixteen delayed claims. For the prespecified response-scale summary, `emmeans` averages the three label levels equally on the logit scale, fixes numeric covariates at their means and random effects at zero, then applies the inverse logit. These are *conditional reference-grid* estimates, not population-average accuracies. The delayed set contains six true, eight false, and two misleading claims; weighting and the order of transformation therefore distinguish the reference-grid summary from raw accuracy. We retain the odds ratio as the primary contrast and label the reference-grid differences explicitly.

Two post-hoc sensitivity summaries check dependence on that calculation. A baseline-adjusted participant-level linear model estimates differences in individual mean accuracy, using its original ordinary model-based interval. A second calculation averages probabilities from the same fitted mixed model over observed delayed covariates and label frequencies, with random effects fixed at zero and fixed-effect delta-method intervals. It changes averaging and transformation, not the fitted model; it does not integrate over random-effect distributions. These summaries are separately labeled and are not independent replications.

Behavioral timing analysis. We calculated active time by summing the initial response time and post-reveal revision time across trials. We report medians with bootstrapped 95% confidence intervals. Condition differences were tested using linear models of log-transformed time, controlling for baseline accuracy. We also used a logistic mixed-effects model to examine whether Wave 1 practice time was associated with delayed accuracy. The model controlled for condition, baseline accuracy, claim truth, and trial order, with random intercepts for participants and claims.

Hypothesis testing. Null-hypothesis tests were two-sided, with $\alpha = .05$. In addition, we performed exploratory equivalence tests, which required the entire 90% interval to fall within that band [17].

4.8 Use of generative AI in the research workflow

Generative AI tools were used to draft wording for the study interface and participant instructions, which the authors reviewed and edited before deployment; to draft portions of the analysis and figure-generation code, which the authors executed on the frozen data and checked against regenerated outputs; and to propose sensitivity and consistency checks, whose appropriateness and interpretation were evaluated by the authors. OpenAI Codex was used during the current revision to compare the documented protocol, deployed implementation, participant logs, analysis code, and reported results. Generative AI did not assign reference labels, interact with participants, access raw participant identifiers, or make final analytical or interpretive decisions. The experimental "Verification Assistant" was a deterministic scripted interface and did not call a generative model during participation.

4.9 Ethics and openness

The study was registered with the authors' institution for research and data-protection compliance on 17 August 2026. Under the requirements applicable in the authors' research environment, this minimal-risk, non-medical behavioural study did not require independent ethics-board review. The record specifies processing ordinary personal data for research under GDPR Article 6(1)(e), with informed consent, and no collection of participants' health data. Before consenting, participants were told that some claims were false or misleading, that nothing in the study was medical advice, and that the full correction sheet would come after the follow-up. Individual participant records were processed and stored pseudonymously.

Supplementary materials are available at [1]. The repository contains the analysis code, participant records, prevalidation materials documenting independent review and adjudication of the experimental claims, a runnable version of the participant-facing web prototype, and README files for reproducing the analyses and running the prototype.

5 Results

5.1 Sample and follow-up

All 552 Wave 1 completers entered the analysis: 184 answer-first, 183 evidence-first, and 185 source-practice. Delayed returners numbered 157, 159, and 158, respectively (474 total; 86% retention). No interface difference in return rates was detected ($\chi^2(2) = 0.22$, $p = .894$, adjusting for baseline accuracy). Returners and non-returners had similar baseline accuracy (0.670 vs. 0.658); returners were somewhat older (41.9 vs. 38.6 years). Baseline accuracy was 0.657, 0.675, and 0.673 across the three groups ($F(2, 549) = 0.52$, $p = .593$). Follow-up occurred after a median of eight days; one returner was marginally outside the planned 6–10-day window (10.001 days).

5.2 Primary comparison: delayed verification

The delayed comparison did not support a five-point evidence-first advantage on the prespecified model summary. For sixteen new claims with evidence cards but no AI answer, observed accuracy was 67.2% for evidence-first, 70.0% for answer-first, and 68.4% for source-practice (Figure 3). Hence, although performance for evidence-first was lower than for answer-first contrast, the difference was not significant: OR = 0.791 (95% CI [0.624, 1.002]), $p = .052$. Thus, [H1] was not supported.

The accompanying conditional reference-grid difference was −5.47 percentage points (95% CI [−10.97, +0.04]; Table 2). The interval includes zero and excludes a +5-point advantage on this specified scale. It also remains compatible with appreciably lower accuracy under evidence-first. Accordingly, the prespecified equivalence test was inconclusive: its 90% interval, [−10.08, −0.85], was not contained within ±5 points. The result constrains the proposed benefit without establishing that the sequences are interchangeable.

5.3 Sensitivity to exclusions, timing, and averaging

The planned exclusions sensitivity, removing the instruction-check failure and reported outside help (no participant met the speed threshold), would lead to OR = 0.79, $p = .051$. Restricting follow-up to the

[1] https://osf.io/j3tce/overview?view_only=780c9262a0c345dabebc9dc506d8da60

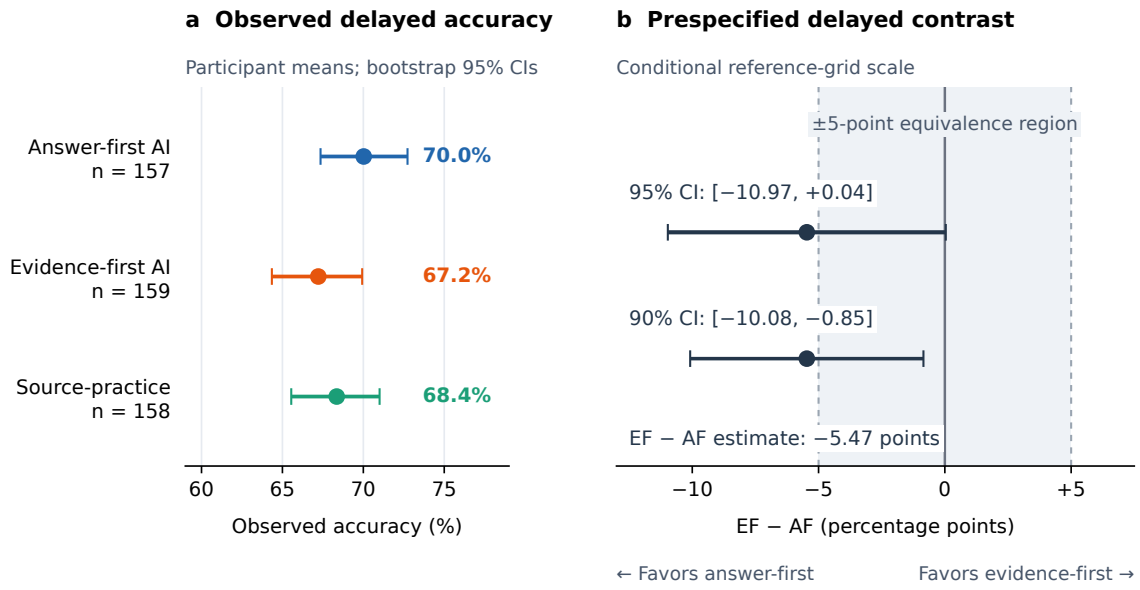


Fig. 3. Delayed verification after a median of eight days. (a) Observed accuracy on sixteen novel claims among returners, with participant-bootstrap 95% intervals over the fixed claim set; these are descriptive. (b) The primary conditional reference-grid evidence-first minus answer-first difference. The 95% interval includes zero but excludes a five-point advantage. The 90% interval is not inside the prespecified ±5-point equivalence band. The two panels use different summaries: observed item frequencies in (a), and equal label weighting on the logit scale, mean covariates, and zero random effects in (b).

planned 6–10-day window ($N = 473$) would lead to OR = 0.79, $p = .052$. These results retain the direction of the original estimate; neither resolves selective attrition.

The participant-level post hoc estimate was −3.68 points (95% CI [−7.43, +0.07], $p = .054$, $d = -0.22$). Averaging fitted probabilities over observed covariates and label frequencies, with random effects zero, gave −3.72 points (95% CI [−7.49, +0.04]; Table 2). The difference in magnitude from the reference grid shows why its −5.47 points should not be presented as the raw or population-average accuracy gap. Across these specified summaries, a near-zero difference remains compatible with the data, the point estimate favors

Table 2. Delayed evidence-first minus answer-first differences. All values are percentage points with 95% intervals. The first row is the prespecified conditional reference-grid summary accompanying the primary OR. The other rows are post hoc sensitivity summaries of the same observations; their averaging and covariance assumptions differ. They are not independent confirmations or interchangeable population effects.

Summary	Difference	95% CI
Prespecified conditional reference grid	−5.47	[−10.97, +0.04]
Participant mean accuracy, baseline adjusted	−3.68	[−7.43, +0.07]
Observed-covariate probability averaging, random effects zero	−3.72	[−7.49, +0.04]

answer-first, and a five-point advantage is not supported. The protocol's equivalence criterion applies to its specified reference-grid calculation.

Delayed performance was especially low on the two misleading claims (21–27% correct across interfaces), compared with true (86–88%) and false (65–69%) claims. Exploratory label-removal checks gave OR = 0.78 [0.60, 1.02] after excluding the misleading item relabeled during validation, and OR = 0.77 [0.56, 1.06] after excluding both misleading items. The point estimates remain similar, but intervals widen. These checks do not replace replication across additional claims; their full results and label-confusion patterns are in the supplement.

5.4 Active comparison and planned secondary outcomes

Source-practice contextualizes the AI sequence comparison, but is not a no-practice control. Its delayed contrast with answer-first was OR = 0.88, 95% CI [0.70, 1.12], $p = .295$; the conditional reference-grid difference was −2.91 points [−8.34, +2.53]. On the separately labeled participant-accuracy scale, evidence-first minus source-practice was −1.73 points [−5.47, +2.01]. These comparisons do not establish equivalence or absolute improvement in any group.

For H2, both AI groups judged the unaided immediate trials at the same positions (5–8), so this comparison is randomized. It favored answer-first: raw accuracy was 69.2% versus 64.3%, OR = 0.71 [0.52, 0.96], unadjusted $p = .028$; the reference-grid difference was −8.2 points [−15.5, −0.8]. Minutes after practicing with an interface that had required them to judge from the evidence before every answer, evidence-first participants were less accurate at judging new claims from evidence than participants who had been handed the answer first. This is a secondary result about the full assigned history up to that point; its Holm-adjusted $p = .055$ warrants caution, and it is not a primary finding of harm. Source-practice's immediate benchmark pools all eight positions (67.6%).

H3's intended within-person removal effect was not found in our analyses. The raw early minus late gaps were 19.4 points in answer-first and 27.7 in evidence-first, opposite to the predicted ordering; the archived block interaction (OR = 2.15 [1.09, 4.22], $p = .029$) is a description of ordered blocks, and its significance does not remove the confound. The supplement preserves the full model and its limits.

5.5 Position-effect analysis (post hoc)

Furthermore, we examined whether the position of a claim in the study affected the results across conditions. Source-practice took the same eight immediate-test positions with no AI compared to the AI conditions, allowing us to compare its wave 1-minus-wave 2 difference with regard to claim position, elapsed time, and the shuffled claim mix alone. That difference was +4.1 points (95% CI [−0.5, +8.6], cluster-robust by participant); the corresponding first-half minus second-half differences were +2.2 points at baseline and −0.1 points at the delayed test, where no AI appeared either (Figure 4). Subtracting the source-practice gap from each AI group's gap (a difference-in-differences) gives +15.4 points [+9.3, +21.5] for answer-first and +23.7 points [+17.7, +29.7] for evidence-first, and the evidence-first gap exceeded the answer-first gap by 8.3 points [2.6, 14.0]. This suggested that the assisted-versus-unaided gaps are far too large to be position effects, and it locates the immediate test in the same pattern as the delayed one: high accuracy while the answer was on screen, and accuracy near the no-AI level as soon as it was not.

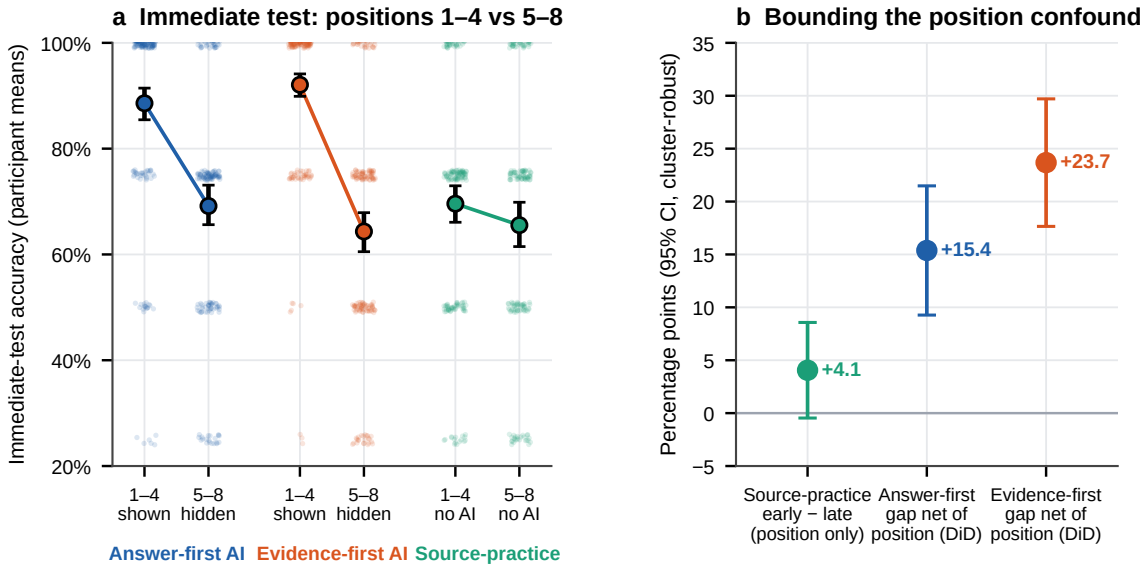


Fig. 4. Bounding the position confound in the immediate test (post hoc). (a) Accuracy on positions 1–4 versus 5–8 by interface (participant means, bootstrap 95% intervals). Both AI groups had the answer on 1–4 and not on 5–8; source-practice had no AI at any position. (b) The source-practice early-minus-late difference estimates the position effect alone; the difference-in-differences subtracts it from each AI group's gap (cluster-robust 95% intervals). The analysis assumes equal position effects across groups and is a bound, not a randomized removal contrast.

We further explored whether performance and choice in Wave 1 predicted outcomes in Wave 2. A 10-point higher rate of selecting the authoritative source during practice was associated with 2.3 points higher delayed accuracy after adjustment for baseline accuracy, baseline source choice and interface (95% CI [1.5, 3.1]); residual confounding and measurement error prevent a causal interpretation. The questionnaire ran the other way from the accuracy results. Evidence-first participants agreed more that they had formed their own judgment before relying on the system (6.19 vs. 5.79 on a 1–7 scale, $d = 0.35$, $p = .001$)—as the interface had required, so the item partly restates the manipulation—and also agreed more that they had gone along with the assistant even when unsure it was right (4.02 vs. 3.29, $d = 0.38$, $p < .001$). They answered about 12% faster on unaided trials (log response time −0.12, 95% CI [−0.21, −0.04]). Reported effort, the feeling that the interface did the thinking, and willingness to use the assistant again did not differ. These items are not validated agency measures; they show that the interface that made people commit first left them feeling more in charge and more deferential at once, while their unaided accuracy was lower. The supplementary analyses report these outcomes, response times, revisions, and questionnaires in full.

5.6 Behavioral timing analysis

In Wave 1, source-practice required less active time than answer-first. In Wave 2, evidence-first required less active time than answer-first (table 3). The other condition differences were not significant. Longer practice time in Wave 1 had a small positive association with delayed accuracy (Spearman's $\rho = .16$, $p < .001$). The adjusted mixed-effects model showed a similar result. A one-standard-deviation increase in log practice time

Table 3. Active task time by condition. Medians are reported with bootstrapped 95% confidence intervals. Adjusted differences were estimated using linear models of log-transformed time with answer-first as the reference condition. Negative values indicate less active time than answer-first.

Wave	Condition	N	Median minutes (95% CI)	Adjusted difference vs. answer-first (95% CI)
Wave 1	Answer-first	184	14.46 [13.73, 15.66]	Reference
Wave 1	Evidence-first	183	15.33 [14.38, 16.87]	6.0% [−3.0%, 15.9%], $p = .194$
Wave 1	Source-practice	185	13.34 [12.27, 14.10]	−8.6% [−16.3%, −0.1%], $p = .046$
Wave 2	Answer-first	157	9.33 [8.48, 10.11]	Reference
Wave 2	Evidence-first	159	7.82 [7.37, 8.90]	−10.9% [−19.5%, −1.4%], $p = .026$
Wave 2	Source-practice	158	9.06 [8.48, 9.58]	−3.1% [−12.4%, 7.3%], $p = .544$

was associated with higher odds of a correct delayed response (OR = 1.25, 95% CI [1.07, 1.46], $p = .004$). The association did not differ significantly between conditions, $ps > .41$.

6 Discussion

We have investigated whether making people decide before a correct AI answer leaves them better at judging new claims once the AI is gone. The results suggest that this is not the case, as the evidence-first sequence has produces no delayed advantage over answer-first. Overall, we have detected no difference between either AI interface, nor between the AI interfaces and source-practice. Below, we first explain what the data suggest, then why our hypotheses have not been supported, and then what it implies for further design and evaluation.

6.1 Verification on loan

Every unaided measurement in this study sits at roughly two-thirds correct: the baseline, participants' own judgments before the answer during practice, the unaided immediate trials, and the delayed test, in every group. Every measurement with the answer on screen sits near 90%. The immediate-test part of that picture cannot be read as a randomized removal effect, because the answer was always shown first; but the no-AI group on the same schedule moved by four points between the same positions, so gaps of 19 and 28 points are not order effects. Descriptively, then, the assistant lent accuracy for the duration of a trial and took it back when it left, and ten practice claims with correct feedback—delivered before the judgment, after it, or as an answer key without AI—did not move unaided accuracy off its starting level. That is the pattern we call verification on loan. It is consistent with studies in which performance with AI and skill without it come apart [2, 12, 31], and with the finding that a correct AI fact-check need not improve discernment [9]. What this study adds is the interface comparison: holding the content of the help fixed and requiring the person to commit before receiving it did not change what remained.

6.2 Why did deciding first not help?

Correction by replacement. Learning research predicts that attempting an answer before being told helps, even when the attempt fails [5, 15]. That benefit is usually measured on the same item; our tests used new claims, so what had to carry over was a way of reading evidence, not an answer. After the reveal, evidence-first participants adopted the AI's answer on 93% of practice trials and moved on; nothing asked them

to reconcile their reasoning with the AI's, which is the step Buçinca et al. found necessary for unaided skill to improve [4]. Committing first created a disagreement; the interface resolved it by replacement. A follow-up would hold the sequence fixed and randomize a reveal that requires the participant to explain the disagreement.

The answer-first explanation as a worked example. Answer-first participants read a correct answer and its explanation before looking at the cards. For beginners, a worked solution studied before attempting problems can teach the pattern better than unaided attempts [33]. Repeated answer exposure could contribute in the same direction. Because the sequences differ in several components, this study cannot separate these; a follow-up should vary the explanation and the exposure independently.

Learning to defer. On about a third of assisted trials, evidence-first participants saw their own judgment fail against an assistant that was never wrong. Afterwards they said more often that they went along with the AI when unsure, said slightly less that the interface made them inspect evidence, and answered faster when the answer was absent. One reading is that repeated correction by an infallible assistant taught them that their reading of the evidence was the less reliable one. This is speculative; it predicts that the evidence-first shortfall would shrink or reverse with an assistant that is visibly wrong on some trials.

Too little practice. All three groups practiced on ten claims in one sitting, and none improved. The study cannot distinguish an interface that fails to teach from a dose too small for anything to teach. Multi-session practice with the same unaided tests would.

6.3 Feeling in control is not being capable

The interface that made people commit first produced the largest questionnaire effect in the study—participants said they had formed their own judgment before relying on the system—and that feeling had no relation to their delayed accuracy. The same participants reported more deference and were less accurate without the AI. Liao and Sundar [20] found that verification actions changed trust without the expected change in felt control ; our results extend the gap to capability. An evaluation that used ownership, effort, or satisfaction as its outcome would have called evidence-first the better design. An evaluation that used accuracy with the AI would have called it marginally better still. Only the unaided tests showed that it built no more skill than the alternative, and that the people who felt most in charge were slightly worse on their own.

6.4 Implications for design and evaluation

Answer timing is not a learning intervention. Requiring a first judgment before the AI's answer remains a reasonable way to reduce acceptance of wrong suggestions in the moment [6, 8]. On this evidence, designers should not expect it to protect or build the user's own skill, and a product that "keeps the human in the loop" by delaying the answer is making a claim about the assisted session, not about what is left afterwards.

If skill is the goal, change what the AI gives. The interventions that have produced unaided gains changed content—hints instead of answers [2], explanations aimed at the user's likely mistake [4], reasoning from an explanation without a recommendation [12]. The evidence-first reveal screen is the natural place to ask the user to explain a disagreement before showing the resolution.

Measure what remains, and audit the schedule. Eight unaided claims in the session and sixteen a week later were enough to separate a large in-session lift from a capability effect bounded within a few points, and the same schedule showed where our own implementation departed from its protocol. A no-AI arm on an identical schedule is cheap and, as here, is what makes a deviation interpretable rather than fatal. We recommend both, and a scripted check, before any model is fitted, that every randomized factor actually varied in the logged data.

6.5 Limitations and scope

The immediate probe did not implement the planned availability randomization. All 367 AI participants received assisted positions 1–4 followed by unaided positions 5–8, although claims were shuffled. Availability is therefore inseparable from order and accumulated exposure; neither a block interaction nor adjustment for position identifies a causal removal cost or differential dependency. The supplementary analyses retain the original estimates with this limitation. The delayed randomized comparison remains a comparison of complete assigned histories, including each interface's four assisted probe trials, rather than an effect of the ten practice claims alone.

All three groups practiced and received reference content. Source-practice supplies an active comparator without AI framing, not a no-practice counterfactual. The study therefore cannot determine whether any group learned more than it would have without practice. Different fixed claim sets across phases likewise prevent interpreting a trajectory as absolute improvement or skill loss. A future study targeting absolute learning should include a no-practice group and counterbalance claim sets across stages.

The outcome is verification using curated evidence. Participants had three prepared cards for each delayed claim, rather than locating and assessing sources on the open web. Only two of the 16 delayed claims were misleading, limiting conclusions about that category. The correct scripted assistant removes model variability but does not represent a fallible or interactive generative system. Ten practice claims and one follow-up also leave longer, spaced practice and other intervals untested.

The sample consisted of U.S. Prolific participants. Although 85.9% returned and retention did not detectably differ by interface, balanced return rates do not exclude selective missingness; the primary analysis uses observed delayed outcomes. The study was not externally preregistered. The team's records describe the hypotheses, specifications, and five-point threshold as fixed before recruitment, but the audit cannot independently verify their timing. The planned power simulation was completed only after data collection. We therefore base the interpretation on estimated effects and intervals, preserve the planning and implementation record, and treat secondary and exploratory findings as supporting context.

7 Conclusion

AI assistants make people better at judging claims while they are present. This study asked whether a common interface choice—hiding the AI's answer until the user has looked at the evidence and committed to a judgment—changes what people can do once the assistant is gone. In a two-wave randomized experiment with identical evidence and an always-correct assistant, it did not: one week later, deciding first produced no advantage and possibly a small disadvantage, neither AI interface differed from practice without AI, and the people who felt most in charge were slightly worse on their own. Accuracy was on loan under both designs. The immediate test did not run as its protocol specified; we report what that removes and what it

leaves, and we treat it as one more argument for the practice the paper recommends: evaluate AI assistance by testing people without it, on new tasks, later, and check that the schedule you analyze is the one you ran.

Acknowledgments

This research received support from the Research Council of Norway (project no. 309339) and industry partners through the Centres for Research-based Innovation scheme, which funds MediaFutures: Research Centre for Responsible Media Technology and Innovation.

Disclosure of AI assistance

Generative AI tools were used to assist with drafting and editing the manuscript, writing and revising code, and analysing the data. The authors reviewed and revised the AI-assisted text and checked the resulting code and outputs. The authors take full responsibility for the manuscript, analyses, and conclusions.

References

[1] Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, et al. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence* 6, 8 (2024), 852–863.

[2] Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakcı, and Rei Mariman. 2025. Generative AI without Guardrails Can Harm Learning: Evidence from High School Mathematics. *Proceedings of the National Academy of Sciences* 122, 26 (2025), e2422633122. doi:10.1073/pnas.2422633122

[3] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48. doi:10.18637/jss.v067.i01

[4] Zana Buçinca, Siddharth Swaroop, Amanda E. Paluch, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2025. Contrastive Explanations That Anticipate Human Misconceptions Can Improve Human Decision-Making Skills. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Article 1024, 25 pages. doi:10.1145/3706598.3713229

[5] Andrew C. Butler, Lisa K. Fazio, and Elizabeth J. Marsh. 2011. The Hypercorrection Effect Persists over a Week, but High-Confidence Errors Return. *Psychonomic Bulletin & Review* 18, 6 (2011), 1238–1244. doi:10.3758/s13423-011-0173-y

[6] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1, Article 188 (2021), 21 pages. doi:10.1145/3449287

[7] Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46. doi:10.1177/001316446002000104

[8] Sander de Jong, Ville Paananen, Benjamin Tag, and Niels van Berkel. 2025. Cognitive Forcing for Better Decision-Making: Reducing Overreliance on AI Systems Through Partial Explanations. *Proceedings of the ACM on Human-Computer Interaction* 9, 2, Article CSCW048 (2025). doi:10.1145/3710946

[9] Matthew R. DeVerna, Harry Yaojun Yan, Kai-Cheng Yang, and Filippo Menczer. 2024. Fact-Checking Information from Large Language Models Can Decrease Headline Discernment. *Proceedings of the National Academy of Sciences* 121, 50 (2024), e2322823121. doi:10.1073/pnas.2322823121

[10] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2015. Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114–126. doi:10.1037/xge0000033

[11] Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Michael Fitzke, Mark Parkinson, Diane Wilson, Paul Fisher, Eric Horvitz, Kori Inkpen, and Besmira Nushi. 2022. Who Goes First? Influences of Human-AI Workflow on Decision Making in Clinical Imaging. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM. doi:10.1145/3531146.3533193

[12] Krzysztof Z. Gajos and Lena Mamykina. 2022. Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*. ACM,

794–806. doi:10.1145/3490099.3511138

[13] Mohammad Hassany, Peter Brusilovsky, Jiaze Ke, Kamil Akhuseyinoglu, and Arun Balajiee Lekshmi Narayanan. 2024. Human-AI co-creation of worked examples for programming classes. *arXiv preprint arXiv:2402.16235* (2024).

[14] Sheer Karny, Lukas William Mayer, Jackie Ayoub, Miao Song, Haotian Su, Danyang Tian, Ehsan Moradi-Pari, and Mark Steyvers. 2024. Learning with AI Assistance: A Path to Better Task Performance or Dependence?. In *Proceedings of the ACM Collective Intelligence Conference*. ACM, 8 pages. doi:10.1145/3643562.3672610

[15] Nate Kornell, Matthew Jensen Hays, and Robert A. Bjork. 2009. Unsuccessful Retrieval Attempts Enhance Subsequent Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 35, 4 (2009), 989–998. doi:10.1037/a0015729

[16] Neema Kotonya and Francesca Toni. 2020. Explainable Automated Fact-Checking for Public Health Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 7740–7754. doi:10.18653/v1/2020.emnlp-main.623

[17] Daniël Lakens, Anne M. Scheel, and Peder M. Isager. 2018. Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science* 1, 2 (2018), 259–269. doi:10.1177/2515245918770963

[18] Danielle Ka Lai Lee and Onur Ramazan. 2021. Fact-checking of health information: The effect of media literacy, metacognition and health information exposure. *Journal of Health Communication* 26, 7 (2021), 491–500.

[19] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (2004), 50–80. doi:10.1518/hfes.46.1.50_30392

[20] Mengqi Liao and S. Shyam Sundar. 2026. Chat but Verify: Combating Misinformation in Conversational Generative AI with Verification Affordance. *Journal of Computer-Mediated Communication* 31, 3, Article zmag012 (2026). doi:10.1093/jcmc/zmag012

[21] Grace Liu, Brian Christian, Tsvetomira Dumbalska, Michiel A. Bakker, and Rachit Dubey. 2026. AI Assistance Reduces Persistence and Hurts Independent Performance. arXiv:2604.04721 [cs.AI] doi:10.48550/arXiv.2604.04721

[22] Jennifer M. Logg, Julia A. Minson, and Don A. Moore. 2019. Algorithm Appreciation: People Prefer Algorithmic to Human Judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103. doi:10.1016/j.obhdp.2018.12.005

[23] Miriam J. Metzger. 2007. Making Sense of Credibility on the Web: Models for Evaluating Online Information and Recommendations for Future Research. *Journal of the American Society for Information Science and Technology* 58, 13 (2007), 2078–2091. doi:10.1002/asi.20672

[24] Zhenni Ni, Yuxing Qian, Shuaipu Chen, Marie-Christine Jaulent, and Cedric Bousquet. 2024. Scientific evidence and specific context: leveraging large language models for health fact-checking. *Online Information Review* 48, 7 (2024), 1488–1514.

[25] Raja Parasuraman and Victor Riley. 1997. Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors* 39, 2 (1997), 230–253. doi:10.1518/001872097778543886

[26] Gordon Pennycook and David G. Rand. 2021. The Psychology of Fake News. *Trends in Cognitive Sciences* 25, 5 (2021), 388–402. doi:10.1016/j.tics.2021.02.007

[27] Griffin Pitts, Muntasir Hoq, Peter Brusilovsky, Narges Norouzi, Arto Hellas, Juho Leinonen, and Bita Akram. 2026. Personalized Worked Example Generation from Student Code Submissions Using Pattern-based Knowledge Components. In *Proceedings of the Thirteenth ACM Conference on Learning@ Scale*. 501–505.

[28] Anku Rani, Valdemar Danry, Paul Pu Liang, Andrew Lippman, and Pattie Maes. 2026. Dialogues with AI reduce beliefs in misinformation but build no lasting discernment skills. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. Article 792, 26 pages. doi:10.1145/3772318.3790656

[29] Rolf Schwonke, Alexander Renkl, Carmen Krieg, Jörg Wittwer, Vincent Aleven, and Ron Salden. 2009. The worked-example effect: Not an artefact of lousy control conditions. *Computers in human behavior* 25, 2 (2009), 258–266.

[30] Judy Hanwen Shen and Alex Tamkin. 2026. How AI Impacts Skill Formation. arXiv:2601.20245 [cs.CY] doi:10.48550/arXiv.2601.20245

[31] David Strömberg, Victor Lei, and Yanhui Wu. 2026. *The Generative AI Learning Penalty: Evidence from Chinese Secondary Education*. CEPR Discussion Paper 21577. CEPR Press, Paris and London. https://cepr.org/publications/dp21577

[32] S. Shyam Sundar. 2020. Rise of Machine Agency: A Framework for Studying the Psychology of Human–AI Interaction (HAII). *Journal of Computer-Mediated Communication* 25, 1 (2020), 74–88. doi:10.1093/jcmc/zmz026

[33] John Sweller and Graham A. Cooper. 1985. The Use of Worked Examples as a Substitute for Problem Solving in Learning Algebra. *Cognition and Instruction* 2, 1 (1985), 59–89. doi:10.1207/s1532690xci0201_3

[34] Harry Yizhou Tian, Hasan Amin, and Ming Yin. 2026. Understanding the Effects of AI-Assisted Critical Thinking on Human-AI Decision Making. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. ACM. doi:10.1145/3772318.3790785

[35] Christoph Trattner. 2026. Epistemic Transfer in AI-Assisted Verification: A Framework and Evaluation Protocol. arXiv:2608.08882 [cs.HC] doi:10.48550/arXiv.2608.08882

[36] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1, Article 129 (2023). doi:10.1145/3579605

[37] Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024. HealthFC: Verifying Health Claims with Evidence-Based Medical Fact-Checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. ELRA and ICCL, 8095–8107. https://aclanthology.org/2024.lrec-main.709/

[38] Sam Wineburg and Sarah McGrew. 2019. Lateral Reading and the Nature of Expertise: Reading Less and Learning More When Evaluating Digital Information. *Teachers College Record* 121, 11 (2019), 1–40.

[39] Shang Wu, Catarina G. Belem, Shuyuan Fu, Mark Steyvers, and Padhraic Smyth. 2026. How AI Assistance Affects Human Skill Development: A Study of Learning with Logic Puzzles. arXiv:2608.23543 [cs.AI] doi:10.48550/arXiv.2608.23543

[40] Haoning Xue and Laramie Taylor. 2023. When do people believe, check, and share health rumors on social media? Effects of evidence type, health literacy, and health knowledge. *Journal of health psychology* 28, 7 (2023), 607–619.