Deciding before AI advice: Greater perceived ownership, no later advantage in health-claim verification

Christoph Trattner[1], Tobias Jovall Wessel[1], Yelyzaveta Lysova[1] and Alain Starke[1]

[1]SFI MediaFutures, University of Bergen, Bergen, Norway

Abstract

Asking people to judge for themselves before they see an AI answer is a common proposal for keeping users in control of AI-assisted decisions. We tested whether this sequence also leaves people better able to judge without the AI afterwards. In a two-session online experiment on health-claim verification, 552 participants practised ten claims with one of three interfaces: an AI answer shown before their judgement (answer-first), the same answer shown only after they had examined the evidence and committed to a judgement (evidence-first), or the same task without AI (source practice). About eight days later, 474 of them judged 16 new claims from evidence cards without an AI answer. Evidence-first participants agreed more strongly that they had formed their own judgement, and also that they had gone along with the assistant when unsure. They were not more accurate a week later: 67.2% of their judgements were correct, against 70.0% after answer-first and 68.4% after the source practice baseline (evidence-first versus answer-first: OR = 0.791, 95% CI [0.624, 1.002], $p = .052$). The data do not support a delayed advantage of evidence-first. The assistant was scripted and always correct and practice was brief, so the result concerns the interaction sequence, not fallible LLM advice or long-term use. We conclude that agency-preserving interfaces need to be evaluated on three separate outcomes: what the interface makes users do, what users report, and what they can do afterwards.

Keywords

Human–AI interaction, large language models, AI advice, cognitive forcing, user agency, health-claim verification, responsible interface design

1. Introduction

Many proposals for agency-preserving AI interfaces share one idea: the user should think first. Instead of presenting an answer that can be accepted with one click, the interface withholds the AI output until the person has looked at the material and committed to a judgement. Such cognitive forcing reduces overreliance on incorrect AI suggestions at the moment of decision [1], and the order in which human and AI judgements are made changes decisions in clinical imaging and diagnosis [2, 3]. For assistants built on large language models (LLMs), which answer fluently and at once, the idea is attractive because it promises to keep the person, and not the model, at the start of the reasoning.

Whether thinking first also leaves people better prepared for the moment when no assistant is available is a different question. Learning research separates performance while support is present from what a person can do afterwards [4], and the difference matters for health information in particular. People ask AI systems about health claims, but they meet most such claims without an assistant at hand, and the answer alone does not guarantee a good decision: in a randomised study with 1,298 participants, people assisted by LLMs assessed medical scenarios no better than people using their usual sources, although the models performed well on their own [5]. Help in the current decision and later independent ability can also come apart. Recommendations with explanations improved nutrition decisions without a learning benefit, whether or not people chose before seeing them [6], and a month of AI-supported news assessment was followed by lower unaided discernment [7]. What has not been tested is a direct,

HUMANE-LLM 2027: Workshop on Human-Centered and Responsible LLMs for Agency-Preserving Intelligent User Interfaces, February 8, 2027, Helsinki, Finland

Christoph.Trattner@uib.no (C. Trattner)



© 2027 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

randomised comparison of the two sequences with identical AI content, with performance measured on new material after a delay and alongside what users did and reported.

We ran this comparison as a two-session experiment with 552 participants. In the first session, participants practised verifying ten health claims with one of three interfaces (Figure 1). In answer-first, a scripted assistant showed its answer and explanation before the participant judged the claim. In evidence-first, participants examined the evidence, chose a source and committed to a judgement before the same answer and explanation appeared, and could then revise. In source practice, participants did the same task without an assistant and received the same content as an answer key. About eight days later, 474 participants returned and judged 16 claims they had not seen before, with evidence cards but without an AI answer.

Our research question (RQ) was: does an evidence-first interaction, in which users examine the evidence and commit to a judgement before they receive the AI's answer, make them more accurate on new claims judged without AI about one week later than an answer-first interaction that presents the same answer before the judgement? We also examined what participants did during practice and how the two sequences differed in the ownership and deference they reported.

In this study it did not. Delayed accuracy was 67.2% after evidence-first and 70.0% after answer-first, and the confidence interval of the contrast leaves almost no room for an evidence-first advantage while still including no difference. The experience reports point the other way: evidence-first participants agreed more strongly that they had formed their own judgement before relying on the system, and they also agreed more strongly that they had gone along with the assistant even when unsure it was right.

The paper adds three things. First, it isolates the interaction sequence: both AI groups saw identical answers and explanations and differed only in the order of steps, with a group that practised without AI as reference. Second, it observes three layers of the same intervention in one sample, namely the action the interface requires (logged first judgements), the experience users report (ownership and deference) and performance without the AI a week later, and shows that they can diverge. Third, it turns this divergence into a recommendation for how agency-preserving interfaces should be evaluated and into design changes that can be tested. This matters because asking users to decide first is a frequently recommended safeguard for LLM assistants, cheap to build and easy to confirm in interaction logs. Judged by whether users comply or by how they feel about the interaction, it will pass; our data show that such evidence says little about what users can do afterwards, which is the outcome that counts when the assistant is absent or wrong.

Two boundaries of the study should be clear from the start. The assistant was scripted and its answers always agreed with the reference labels, so we isolate the interaction sequence and do not evaluate a live, fallible LLM; and practice consisted of ten claims in one sitting, so the study describes a brief interface experience and not sustained use. The study belongs to a line of work arguing that AI assistance should be judged by what users retain when it is removed [8, 9]; a framework preprint summarises selected outcomes of this cohort [10], and the present paper is the full empirical report.

2. Background

2.1. Deciding first as a way to preserve agency

Appropriate reliance means accepting advice when it is likely to be right and resisting it when it is not [11]. Cognitive forcing functions change what a user has to do before advice can be accepted, and Buçinca et al. showed that they reduce overreliance on incorrect suggestions, at some cost to user experience [1]. The order of the two judgements can also affect accuracy while the advice is present: in two exploratory studies of medical diagnosis, protocols that showed the AI's advice first were more accurate than protocols in which the physician judged first, and the authors nevertheless recommend human-first protocols at intervals to monitor deskilling, which they did not measure [3]. These results concern the assisted decision itself. They leave open what a required first judgement contains, since it can come from careful reading of the evidence, from a guess, or from the wish to get to the next screen, and what it leaves behind.

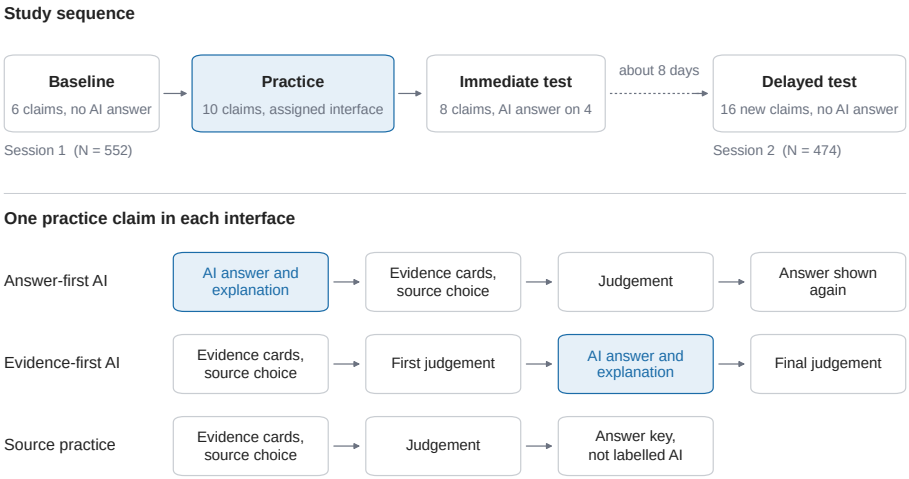


Figure 1: Study sequence (top) and the order of steps for one practice claim in each interface (bottom). The AI answer and explanation (blue) had the same content in both AI interfaces; source practice presented it as an answer key.

Agency is also broader than a required action. Sundar distinguishes the agency of the machine, the affordances of the interface and the psychological response of the user [12], and the layers do not always move together: in an experiment with a health chatbot that gave false information, users who acted on a verification option trusted the chatbot less, while their sense of agency did not differ from that of users without the option [13]. Our experiment covers three observable parts of this picture: an action the interface requires (a judgement before the answer), what participants say about the experience, and how accurately they judge later without the answer.

2.2. Assisted performance and what remains afterwards

Gajos and Mamykina separated assistance that improves the current decision from assistance that supports incidental learning. In three experiments on nutrition decisions, a recommendation with an explanation improved decisions but produced no learning within the session, both when it was shown before the participant's choice and when it was shown after it; only an explanation without a recommendation produced learning [6]. Buçinca et al. later found that explanations which contrast the correct answer with a plausible misconception can improve unaided decisions [14].

Recent studies with LLMs report costs once the assistance is withdrawn. Students who practised mathematics with an unrestricted LLM tutor did worse on a later exam than students who had practised without AI [15]. In logic puzzles, participants with cheap access to an always-correct AI tended to improve less than participants without AI [16]. For misinformation, fact-checks from an LLM did not reliably improve discernment of headlines even while they were shown [17], and a month of AI-supported news assessment produced strong assisted performance but lower unaided discernment on unseen items [7]. Interface design can work against this: feedback that made the consequences of offloading explicit reduced offloading to an LLM and improved performance on an unaided test [18].

Testing people later, on new material and without the AI, is therefore an established approach, and the finding that a prior choice does not by itself produce learning has a precedent. New here are the direct randomised comparison of the two sequences with identical content, the test a week later instead of in the same session, and the logged first judgements and experience reports.

Learning research gives reasons to expect an advantage for either sequence. Attempting an answer before receiving feedback can improve later learning even when the attempt fails [19], which favours evidence-first. Studying a worked solution can help novices more than solving the problem themselves [20], which favours answer-first. Both accounts assume that there is something general to learn. In our task this is the procedure that every trial repeats: identify the most authoritative of several sources and

compare what the claim says with what the evidence says. The delayed claims were new, so remembered answers could not help, and any benefit had to come from this shared procedure.

Transfer of this kind is demanding, because people often fail to apply a procedure to a new case unless they have abstracted it from several examples [21]. We test on new claims because this is the criterion implied when deciding first is recommended as a way to keep people able to judge on their own; Section 6 states what a test on the practised claims would have added.

2.3. Health claims

Judging health information depends on more than the text on the screen. Health literacy, prior knowledge and the personal relevance of a claim all play a role [22, 23]. We chose the domain because checking a health claim with the help of an AI system is an everyday use of decision support. The task gives participants curated evidence cards, so it tests judgement when evidence is at hand and not search on the open web.

3. Method

3.1. Design and procedure

The experiment had three between-subjects conditions and two sessions (Figure 1). In the first session, all participants judged six baseline claims without an AI answer or feedback, then ten practice claims in their assigned interface, then eight claims in an immediate test. About a week later they judged 16 new claims in the delayed test, where no AI answer was shown in any condition. Claims were presented in random order within each stage, and each stage used its own set of claims. Because the sets differ, a change in accuracy from one stage to the next cannot be read as learning or forgetting; the informative comparison is between conditions at the same stage.

With these numbers of claims, the first session took a median of 19.5 minutes and the second 9.8 minutes. The numbers were not derived from a calibration study. The experiment therefore tests a brief exposure to an interface and not a training programme.

3.2. Participants

Participants were recruited on Prolific on 24 and 25 August 2026. They had to be born in and resident in the United States, with an approval rate of at least 95% and at least 100 completed submissions. A total of 552 people completed the first session and were randomly assigned to answer-first ($n = 184$), evidence-first ($n = 183$) or source practice ($n = 185$). All were invited back, and 474 (85.9%) completed the delayed test after a median of 8.0 days: 157, 159 and 158 in the three conditions. Return rates did not differ between conditions ($\chi^2(2) = 0.22$, $p = .894$). Returners and non-returners had similar baseline accuracy (67.0% and 65.8%); returners were somewhat older (41.9 versus 38.6 years). The planned return window was six to ten days; one participant returned just after it had closed.

Among the returners, age ranged from 19 to 80 years ($M = 41.9$, $SD = 12.7$); 246 were women, 223 men, four non-binary, and one preferred to self-describe. Participants received GBP 5.00 for the first session and GBP 2.00 for the second.

3.3. Interfaces

The three interfaces showed the same claims, the same three evidence cards per claim and the same reference answer and explanation. They differed in the order of the steps (Figure 1). In answer-first, the assistant's answer and explanation appeared first; participants then saw the evidence cards, chose the best source, gave their judgement, and saw the answer once more. In evidence-first, participants saw the evidence cards, chose the best source and gave a first judgement; only then did the assistant's answer and explanation appear, after which participants gave a final judgement. In source practice, participants saw

the evidence, chose a source and judged the claim, and then received the correct answer and explanation as an answer key that was not presented as AI.

Two properties of this design matter for interpretation. First, the two AI interfaces differ in more than timing: evidence-first requires two judgements and answer-first shows the answer twice, so we compare complete interaction sequences and cannot attribute a difference to timing alone. Second, the assistant was a deterministic script, not a generative model, and every answer it gave matched the reference label. This removes variation in advice quality between participants, and it means that the study cannot say how either interface performs when the AI is wrong.

3.4. Claims and evidence

The 40 English-language claims covered antibiotics, sleep, physical activity, vitamins and supplements, sunscreen, detox products, colloidal silver, probiotics, and diet and cancer. Each claim was labelled true, false or misleading, where misleading means that the statement contains some truth but invites a wrong conclusion. The claims were written for the study, and each came with three evidence cards: a summary of an authoritative public source (such as the CDC, WHO, FDA, NIH or the National Cancer Institute), a summary of a contextual or limiting source, and an illustrative lower-quality or promotional source.

Two team members independently labelled all 40 claims without seeing the intended labels or the stage to which a claim belonged. They agreed on 33 of 40 claims (82.5%; Cohen's $\kappa = .70$), and all seven disagreements concerned the boundary between false and misleading. These were resolved against the sources, with wording or evidence cards edited where needed. The delayed test contained six true, eight false and two misleading claims, so it says little about misleading claims as a category.

3.5. Measures

On every trial, participants chose the source they considered best, classified the claim as true, false or misleading, and rated their confidence from 0 (guessing) to 100 (certain). The primary outcome is whether the judgement in the delayed test was correct.

Before the task, participants reported age, gender, education, how often they use AI tools, their general trust in AI (1–7), their confidence in spotting false health claims (0–100), and their agreement with five statements about how they check information (1–5; averaged, Cronbach's $\alpha = .78$). These are descriptive self-reports, not validated scales of health literacy. We use them to describe the sample and, in an exploratory analysis, to ask whether the result depends on them.

After the first session, participants rated statements about the experience on 1–7 scales, among them whether they had formed their own judgement before relying on the system's answer. Participants in the two AI conditions also rated the statement "I went along with the assistant even when I was unsure it was right." We refer to these two items as perceived ownership and reported deference. Both are single items written for this study. The ownership item partly restates what evidence-first required, which is a reason to read it as a report of the experience and not as a measure of agency.

3.6. Analysis

The primary analysis includes the 474 participants with delayed outcomes. A mixed-effects logistic regression predicts the correctness of each delayed judgement from interface, baseline accuracy, claim label and trial position, with random intercepts for participants and claims [24]. The primary contrast is evidence-first versus answer-first; odds ratios below one favour answer-first, and all tests are two-sided. Because odds ratios are hard to read as accuracy, we also express the contrast in percentage points. An equivalence analysis compares the 90% confidence interval of the model-based difference with a margin of ± 5 percentage points [25].

Three sensitivity analyses address data quality: excluding the one participant who failed the instruction check and the eight who reported outside help, restricting the sample to the planned return window, and removing the two misleading claims. The comparison of the two AI conditions on the immediate test is

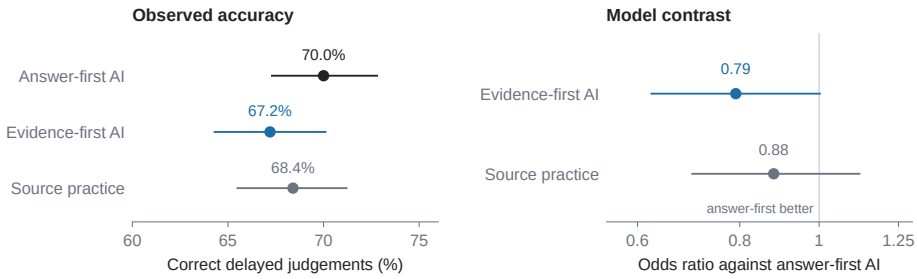


Figure 2: Delayed test, 474 returners. Left: observed accuracy on 16 new claims judged without an AI answer. Right: odds ratios from the primary mixed model against answer-first; values below one mean lower odds of a correct judgement. Bars are 95% confidence intervals.

secondary and is reported with and without Holm adjustment. All analyses of experience reports, practice behaviour and participant characteristics are exploratory.

3.7. Deviations and transparency

The study was not preregistered with an external registry. The hypotheses and model specifications are documented in the team's internal records from before recruitment, which is weaker evidence of prior planning than a public registration. We therefore report estimates with confidence intervals, do not rely on significance alone, and treat the equivalence margin as a benchmark for interpretation. The power simulation was completed only after data collection and is not used to justify the sample size.

One part of the implementation deviated from the plan. In the immediate test, the AI answer was meant to be available on a random half of the eight claims. As implemented, both AI groups received the answer on the first four claims and no answer on the last four. The drop in accuracy from the first to the second half therefore mixes the removal of the answer with the position in the test, and we do not interpret it. The comparison between the two AI groups on the last four claims is unaffected, because both groups judged those claims at the same positions and without an answer.

3.8. Ethics and materials

The study was registered with the authors' institution for research and data-protection compliance on 17 August 2026. According to the institutional record, review by an ethics board was not required for this minimal-risk, non-medical behavioural study. Participants gave informed consent, were told that some claims were false or misleading and that the study did not give medical advice, and were told that a complete correction sheet would follow the delayed test. Data were processed under pseudonyms. Participants were instructed not to use search engines, AI tools or other outside help during the sessions. Materials, the implemented task and analysis documentation are available in an OSF repository.[1]

4. Results

4.1. Delayed accuracy

One week after practice, evidence-first participants were not more accurate than answer-first participants (Figure 2). Observed accuracy on the 16 new claims was 70.0% after answer-first, 67.2% after evidence-first and 68.4% after source practice. The primary contrast between evidence-first and answer-first was OR $= 0.791$, 95% CI [0.624, 1.002], $p = .052$. The estimate lies on the side of answer-first, and the interval reaches just past one. The hypothesis of an evidence-first advantage is not supported.

[1] https://osf.io/j3tce/overview?view˙only=780c9262a0c345dabebc9dc506d8da60

Table 1
Characteristics of the 474 returners by interface (157, 159 and 158 participants), measured before practice. Values are means (SD) or percentages.

	Answer-first	Evidence-first	Source practice
Age in years	40.8 (11.8)	42.7 (12.7)	42.2 (13.4)
Bachelor's degree or higher	61.8%	64.2%	62.0%
Uses AI tools daily or almost daily	52.9%	46.5%	48.7%
General trust in AI (1–7)	4.76 (1.45)	4.70 (1.41)	4.62 (1.42)
Confidence in spotting false health claims (0–100)	62.3 (23.1)	64.4 (22.8)	63.7 (20.8)
Checking habits, five items (1–5)	4.32 (0.64)	4.35 (0.60)	4.31 (0.56)
Baseline accuracy, six claims	65.8%	68.4%	66.7%

Expressed in percentage points, the model-based difference between evidence-first and answer-first is -5.5 (95% CI [-11.0, 0.0]). It is larger than the difference between the observed percentages because it refers to a typical participant and claim and weights the three claim labels equally, although the delayed test contained only two misleading claims. Two post hoc summaries of the same data that keep the observed mix of claims both give -3.7 points (the model averaged over the observed claims and participants: 95% CI [-7.5, 0.0]; participants' mean accuracy adjusted for baseline accuracy: [-7.4, $+0.1$]). All three intervals end close to zero at the upper bound, so an evidence-first advantage of practical size is unlikely, whereas a disadvantage of several points is compatible with the data.

The equivalence analysis leads to the same reading from the other side. The 90% interval of the model-based difference, [-10.08, -0.85], does not lie within ± 5 points, so we cannot conclude that the two interfaces are equivalent. The result is thus neither evidence of harm from evidence-first nor evidence that the sequence makes no difference; it is evidence against the expected benefit.

The estimate was stable across the sensitivity analyses. Excluding participants with quality flags gave OR $= 0.79$ ($p = .051$), and restricting the sample to the planned return window gave OR $= 0.79$ ($p = .052$). Accuracy on the two misleading claims was low in all conditions (21–27%, against 86–88% for true and 65–69% for false claims); without these two claims the contrast was OR $= 0.77$, 95% CI [0.56, 1.06].

Neither AI interface left participants reliably more accurate than practice without AI: source practice did not differ reliably from answer-first (OR $= 0.88$, 95% CI [0.70, 1.12], $p = .295$), and observed accuracy after evidence-first was below that after source practice. Since the source-practice group also practised and received correct answers, it shows how the AI conditions compare with practice without AI, not with no practice.

4.2. Participant characteristics

Health-related judgement varies between people, so we examined whether the three groups differed in characteristics that could matter and whether the result depended on them. Table 1 shows that the groups of returners were similar in baseline accuracy, self-reported checking habits, confidence in spotting false health claims, education, AI use and trust in AI. Baseline accuracy was the strongest predictor of delayed accuracy ($r = .34$), followed by checking habits ($r = .18$). Confidence in spotting false health claims was unrelated to delayed accuracy ($r = .05$).

In exploratory participant-level models, the difference between evidence-first and answer-first did not vary with any of these characteristics (six interaction tests, all $p \geq .21$), and adding all of them as covariates left the estimate at -3.6 points (95% CI [-7.3, $+0.1$]). The study could only have detected large interactions. Within this sample, however, there is no sign that evidence-first helped a particular subgroup, for example participants with careful checking habits or with low initial accuracy.

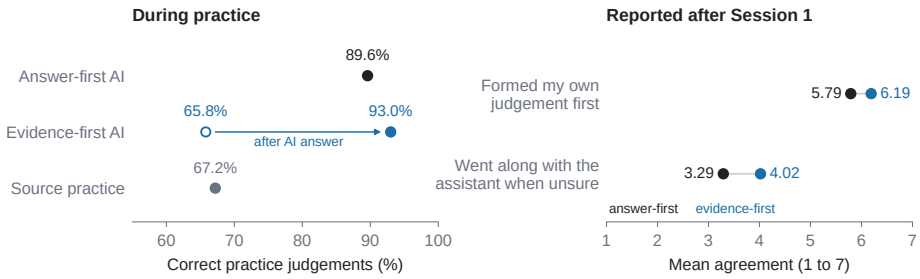


Figure 3: Left: accuracy on the ten practice claims (474 returners); evidence-first participants judged before (open circle) and after (filled circle) the AI answer. Right: mean agreement of the two AI groups with two experience statements after the first session, from 1 to 7 (all 552 participants).

4.3. What happened during practice

The practice data show what the required first judgement consisted of (Figure 3, left). Evidence-first participants were correct in 65.8% of their first judgements. This is the same level as in source practice, where participants judged the same claims without an assistant (67.2%), and close to the baseline level of all groups. Committing to a judgement before an AI answer therefore did not make that judgement more accurate than a judgement made with no AI in prospect. After the assistant's answer appeared, evidence-first participants changed their judgement on 29.1% of trials. They corrected 81% of their wrong first judgements, almost never turned a correct judgement into a wrong one (0.5% of trials), and reached a final accuracy of 93.0%. Answer-first participants, who saw the answer before judging, were correct on 89.6% of practice trials. On roughly one trial in three, evidence-first participants thus saw an independent judgement overruled by an assistant that was never wrong, and the interface then asked them for a new answer, not for the reason why the first had been wrong.

Across conditions, participants who chose the authoritative source more often during practice were more accurate a week later: ten percentage points more authoritative choices went with 2.3 points higher delayed accuracy, adjusted for baseline accuracy, baseline source choice and interface (95% CI [1.5, 3.1]). This is an association between people and may reflect stable differences in care or skill. The interfaces did not differ in this step of the procedure a week later: participants chose the authoritative source on 74.7%, 73.1% and 74.6% of delayed trials after answer-first, evidence-first and source practice (one-way ANOVA on participant means, $p = .79$).

4.4. Reported experience

The experience reports differ from the accuracy results (Figure 3, right). Evidence-first participants agreed more strongly than answer-first participants that they had formed their own judgement before relying on the system (6.19 versus 5.79; $d = 0.35$, $p = .001$). They also agreed more strongly that they had gone along with the assistant even when unsure it was right (4.02 versus 3.29; $d = 0.38$, $p < .001$). Among returners in the two AI conditions, reported deference was negatively related to delayed accuracy ($r = -.22$, $p < .001$), whereas perceived ownership was unrelated to it ($r = .05$). These tests are exploratory and unadjusted for multiple comparisons.

4.5. Immediate test

Minutes after practice, the two AI groups judged four claims without an AI answer. Accuracy was 69.2% after answer-first and 64.3% after evidence-first (OR = 0.71, 95% CI [0.52, 0.96]; $p = .028$ unadjusted, $p = .055$ after Holm adjustment). The direction matches the delayed result; as a secondary outcome that does not survive adjustment, it is supporting information. On the four preceding claims, where the AI answer was shown, accuracy was 88.6% and 92.1%. As explained in Section 3.7, we do not interpret the difference between the two halves of the test.

5. Discussion

5.1. Accurate with the answer, no advantage a week later

With the AI answer on screen, participants in both AI interfaces were right about nine times in ten. A week later, on new claims and without the answer, they were right about two times in three, and neither group was reliably more accurate than the group that had practised without AI. Each stage used different claims, so the drop itself does not measure forgetting. What the comparison with source practice shows is that the help received during practice left no measurable advantage for the next claims, the pattern that learning research describes as performance without learning [4].

The practice data suggest why committing first did not change this, although they cannot prove it. The first judgement in evidence-first was real, in the sense that it was made without knowledge of the answer, but it was no better than a judgement made without any AI. The requirement to commit therefore did not, by itself, lead participants to read the evidence more successfully. What followed the first judgement was also undemanding. When the assistant disagreed, participants could replace their answer with the assistant's, and in most cases they did. Research on learning from errors suggests that a failed attempt helps when the learner then processes the correct solution [19]. Our interface made the attempt mandatory and left the processing optional.

Answer-first, in turn, may have been a better learning condition than its reputation suggests. Participants saw the correct answer and its explanation while they read the evidence cards, which resembles studying a worked example [20]. The estimate in favour of answer-first is consistent with this account, but the study was not designed to test it.

Finally, the pattern of reported deference fits what participants experienced. An assistant that is right every time, and that corrects the user on one trial in three, gives the user good reason to defer; people check an automated aid more closely once they have seen it fail [26], and our participants never did. Going along with such an assistant is not a reasoning failure, and under an always-correct assistant it cannot be called overreliance. It does raise the possibility that deciding first, when followed by reliable correction, teaches people that the assistant reads the evidence better than they do. Whether that lesson carries over to situations in which the assistant is wrong is the question our design cannot answer.

5.2. Ownership is not evidence of capability

Perceived ownership rose in the condition in which later accuracy did not, and within the AI conditions it was unrelated to later accuracy. An evaluation of evidence-first by interaction log would find that every user made an independent judgement, and an evaluation by questionnaire would find that users felt the judgement was their own. Both findings would be correct, and neither would reveal that users were no better prepared for the next claim. A comparable dissociation between what users do and what they report has been observed for verification options in a health chatbot [13], and our result for later learning agrees with the same-session finding of Gajos and Mamykina [6].

For the evaluation of agency-preserving interfaces we therefore suggest keeping three outcomes apart and stating in advance which of them a design is meant to change. The first is the required action: what the interface makes the user do, which logs can verify. The second is the reported experience: whether users feel that the judgement was theirs. The third is subsequent capability: how users judge when the support is absent, changed or wrong [9], which requires a later test, and for errors a test that contains erroneous advice. Success on one of these outcomes should not be reported as success on the others.

5.3. Implications for LLM-based interfaces

The results suggest design changes that can be tested with the sequence held fixed. The first is to require reconciliation: when the user's judgement and the AI answer differ, the interface asks the user to identify the piece of evidence that decides between them before moving on. The second is to change the explanation, for example to one that contrasts the answer with the user's actual first judgement, in line with the benefit of contrastive explanations [14]. The third is to spread practice over several sessions and

to build it around recurring evidential patterns, such as a claim that overstates what a cited study found. A fourth follows from a process account that the first author developed after this study, in which people retain what they check themselves, and check when the task requires it or when they regard the assistant as fallible [27]. On this account, a task that cannot be completed without checking the evidence, or an assistant with a disclosed error rate, should prompt more checking than a first judgement that the answer then replaces; the account has so far been examined in simulations only.

A scripted, correct assistant is a deliberate simplification. Deployed LLMs make errors, express uncertainty unevenly and respond differently to the same question, and each of these properties can change how users treat a first judgement of their own. An evidence-first sequence may prove more valuable when the assistant is sometimes wrong, because the user's first judgement is then the only safeguard. Our data cannot support or refute this. They do show that the benefit should not be assumed from the sequence alone, and that it has to be demonstrated with a test in which the outcome of interest, catching an error or judging without the assistant, is measured directly.

6. Limitations

The exposure was brief: ten practice claims in one sitting and a single delayed test after about a week. The study cannot describe sustained use or the course of an effect over time and, because each stage used different claims and no group went without practice, it estimates differences between interfaces and not how much any group learned. It also contains no test on the practised claims themselves, which would have shown whether the interfaces differ in the retention of specific answers. A follow-up study should combine several practice sessions with later tests on the practised claims and on new claims that share their evidential pattern, and should track whether the way participants use and regard the assistant persists.

The assistant was always correct, which rules out conclusions about error detection and overreliance. The two AI interfaces differed in the number of judgements and answer exposures as well as in order, and the implementation deviation in the immediate test prevents a causal estimate of what happens when the AI answer is withdrawn.

The delayed test had only two misleading claims, the evidence was curated, and the sample was an experienced online panel from the United States. Checking habits and confidence were measured with self-report items written for this study; validated measures of health literacy, prior beliefs about the claims, self-efficacy and medical background were not collected, and neither was exposure to health information or AI tools between the sessions. Random assignment makes such exposure equally likely in all groups, but the interfaces themselves may have changed what participants did in the intervening week. Seventy-eight participants did not return; return rates and baseline accuracy were similar across groups, which limits but does not exclude selective attrition. The study was not externally preregistered, and the experience and moderation analyses are exploratory.

7. Conclusion

Requiring people to judge a health claim before they saw a correct AI answer made them feel that the judgement was their own, and it made them more inclined to go along with the assistant. It did not make them more accurate on new claims a week later; the estimate favoured the interface that showed the answer first, without establishing a reliable difference in either direction. For the design of agency-preserving LLM interfaces, the lesson is that a required judgement is a starting point and not an outcome. Designs that aim to preserve what users can do on their own should say what users are meant to learn, give them the opportunity to practise it, and test it when the assistant is no longer there.

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.

Declaration on Generative AI

Generative AI tools supported the drafting of interface and instruction wording, portions of the analysis and figure code, and consistency checks in the original research workflow, as documented in the study materials. For this workshop manuscript, OpenAI ChatGPT/Codex was used to restructure and edit an earlier draft and to prepare its LaTeX formatting, and Anthropic Claude was used to review and edit the text, to recompute the descriptive statistics and exploratory participant-level analyses from the study data, and to draw the figures. The experimental assistant shown to participants was a deterministic script and did not call a generative model. The authors reviewed and edited all content and take full responsibility for the publication.

References

[1] Z. Buçinca, M. B. Malaya, K. Z. Gajos, 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 (2021). doi:10.1145/3449287.

[2] R. Fogliato, S. Chappidi, M. Lungren, et al., Who goes first? Influences of Human-AI workflow on decision making in clinical imaging, in: Proceedings of ACM FAccT 2022, ACM, 2022. doi:10.1145/3531146.3533193.

[3] F. Cabitza, A. Campagner, L. Ronzio, et al., Rams, hounds and white boxes: Investigating Human–AI collaboration protocols in medical diagnosis, Artificial Intelligence in Medicine 138 (2023) 102506. doi:10.1016/j.artmed.2023.102506.

[4] N. C. Soderstrom, R. A. Bjork, Learning versus performance: An integrative review, Perspectives on Psychological Science 10 (2015) 176–199.

[5] A. M. Bean, R. E. Payne, G. Parsons, et al., Reliability of LLMs as medical assistants for the general public: A randomized preregistered study, Nature Medicine 32 (2026) 609–615. doi:10.1038/s41591-025-04074-y.

[6] K. Z. Gajos, L. Mamykina, Do people engage cognitively with AI? Impact of AI assistance on incidental learning, in: Proceedings of ACM IUI 2022, ACM, 2022, pp. 794–806. doi:10.1145/3490099.3511138.

[7] A. Rani, V. Danry, P. P. Liang, A. Lippman, P. Maes, 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, 2026. doi:10.1145/3772318.3790656.

[8] C. Trattner, Verification on loan: When AI checks facts so we don't have to, Communications of the ACM (2026). doi:10.1145/3861070, accepted for publication.

[9] C. Trattner, Your AI productivity gains may be hiding a capability gap, California Management Review Insights (2026). 12 November 2026.

[10] C. Trattner, Epistemic transfer in AI-assisted verification: A framework and evaluation protocol, 2026. doi:10.48550/arXiv.2608.08882.

[11] J. D. Lee, K. A. See, Trust in automation: Designing for appropriate reliance, Human Factors 46 (2004) 50–80. doi:10.1518/hfes.46.1.50_30392.

[12] S. S. Sundar, Rise of machine agency: A framework for studying the psychology of Human–AI interaction (HAII), Journal of Computer-Mediated Communication 25 (2020) 74–88. doi:10.1093/jcmc/zmz026.

[13] M. Liao, S. S. Sundar, Chat but verify: Combating misinformation in conversational generative AI with verification affordance, Journal of Computer-Mediated Communication 31 (2026). doi:10.1093/jcmc/zmag012.

[14] Z. Buçinca, S. Swaroop, A. E. Paluch, F. Doshi-Velez, K. Z. Gajos, 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, 2025. doi:10.1145/3706598.3713229.

[15] H. Bastani, O. Bastani, A. Sungu, H. Ge, Ö. Kabakcı, R. Mariman, Generative AI without guardrails can harm learning: Evidence from high school mathematics, Proceedings of the National Academy of Sciences 122 (2025) e2422633122. doi:10.1073/pnas.2422633122.

[16] S. Wu, C. G. Belem, S. Fu, M. Steyvers, P. Smyth, How AI assistance affects human skill development: A study of learning with logic puzzles, in: Proceedings of the 2026 ACM Conference on Human-AI Complementarity and Alignment (HCOMP 2026), ACM, 2026. doi:10.1145/3834580.3838741.

[17] M. R. DeVerna, H. Y. Yan, K.-C. Yang, F. Menczer, Fact-checking information from large language models can decrease headline discernment, Proceedings of the National Academy of Sciences 121 (2024) e2322823121. doi:10.1073/pnas.2322823121.

[18] S. Maier, K. Schwabe, M. Schneider, S. Feuerriegel, Designing against deskilling: Metacognitive feedback reduces cognitive offloading to LLM assistants, arXiv preprint arXiv:2609.20143, 2026. doi:10.48550/arXiv.2609.20143.

[19] N. Kornell, M. J. Hays, R. A. Bjork, Unsuccessful retrieval attempts enhance subsequent learning, Journal of Experimental Psychology: Learning, Memory, and Cognition 35 (2009) 989–998. doi:10.1037/a0015729.

[20] J. Sweller, G. A. Cooper, The use of worked examples as a substitute for problem solving in learning algebra, Cognition and Instruction 2 (1985) 59–89. doi:10.1207/s1532690xci0201_3.

[21] M. L. Gick, K. J. Holyoak, Schema induction and analogical transfer, Cognitive Psychology 15 (1983) 1–38.

[22] D. K. L. Lee, O. Ramazan, Fact-checking of health information: The effect of media literacy, metacognition and health information exposure, Journal of Health Communication 26 (2021) 491–500.

[23] H. Xue, L. Taylor, 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 (2023) 607–619.

[24] D. Bates, M. Mächler, B. Bolker, S. Walker, Fitting linear mixed-effects models using lme4, Journal of Statistical Software 67 (2015) 1–48. doi:10.18637/jss.v067.i01.

[25] D. Lakens, A. M. Scheel, P. M. Isager, Equivalence testing for psychological research: A tutorial, Advances in Methods and Practices in Psychological Science 1 (2018) 259–269. doi:10.1177/2515245918770963.

[26] J. E. Bahner, A.-D. Hüper, D. Manzey, Misuse of automated decision aids: Complacency, automation bias and the impact of training experience, International Journal of Human-Computer Studies 66 (2008) 688–699. doi:10.1016/j.ijhcs.2008.06.001.

[27] C. Trattner, Verification learning theory: A process model of checking and retention in AI-assisted judgment, arXiv preprint arXiv:2610.37875, 2026. doi:10.48550/arXiv.2610.37875.