Epistemic Transfer in AI-Assisted Verification: A Framework and Evaluation Protocol

Christoph Trattner
SFI MediaFutures, Research Centre for Responsible Media Technology and Innovation
Department of Information Science and Media Studies, University of Bergen, Norway
christoph.trattner@uib.no • ORCID: 0000-0002-1193-0508

August 2026

Working paper. Comments welcome.

Abstract

AI tools that help people judge online claims are usually evaluated while the tool is present. This paper asks a different question: after using such a tool, what can the user still do on their own? I call this *epistemic transfer*. It refers to the effect of prior AI-assisted verification on later unassisted performance on new claims. In this paper, I make three contributions. First, I distinguish epistemic transfer from nearby outcomes such as correction effects, trust, reliance, and human–AI team performance. Second, I introduce two simple quantities for studying it: the Epistemic Transfer Effect (ETE), which compares delayed unassisted performance across conditions, and Tool-Removal Cost (TRC), which measures the immediate drop in performance when the tool is taken away. Third, I turn these ideas into a practical evaluation protocol that can be used in online experiments or field studies. The protocol combines answer-first and evidence-first AI conditions with active-practice and no-practice controls, delayed tests on held-out claims, behavioral measures, and participant- and item-level analyses. Putting ETE and TRC together yields a diagnostic space that separates capability building, capability plus tool advantage, epistemic inertness or de-skilling, and verification on loan. The point is not that every AI tool must teach. The point is that when independent judgment matters, we should test not only whether a tool helps now, but also what it leaves behind.

Keywords: epistemic transfer; AI-assisted verification; fact-checking; cognitive offloading; de-skilling; human–AI interaction; evaluation methodology

1 Introduction

AI systems increasingly help people judge whether online information is true, misleading, or unreliable. Some systems retrieve evidence and predict claim veracity (Guo et al., 2022). Others summarize sources and produce fluent, natural-language verdicts. Platforms also add labels, provenance signals, or other credibility cues. Most evaluations of these systems focus on what happens at the moment of use: whether the tool is accurate, whether users follow good advice, or whether the human–AI pair performs better than the human or the AI alone (Bansal et al., 2021; Vaccaro et al., 2024).

Those are important outcomes, but they are not the whole story. In many settings, the bigger question is what repeated use of such tools does to the user. A helpful system may expose people

to better evidence and better strategies. But it may also reduce the need to search, compare sources, inspect uncertainty, or form an independent judgment. In other words, the same system can improve performance now while having positive, null, or negative effects on later independent performance.

This issue is familiar in nearby literatures. Learning research distinguishes doing well during supported practice from actually learning something that remains after support is removed (Soderstrom and Bjork, 2015; Roediger III and Karpicke, 2006). Cognitive-offloading research shows that people shift effort and memory when external resources are available (Risko and Gilbert, 2016; Fisher et al., 2015; Storm and Stone, 2015). Automation research shows that high levels of support can reduce opportunities to maintain manual or cognitive skills (Bainbridge, 1983; Casner et al., 2014; Onnasch et al., 2014; Parasuraman and Riley, 1997). Recent AI studies suggest that the same concern matters for generative and clinical AI systems as well (Bastani et al., 2025; Budzyń et al., 2025; DeVerna et al., 2024; Rani et al., 2026; Liu et al., 2026; Shen and Tamkin, 2026).

What is missing is a simple way to study this systematically for AI-assisted verification. This paper aims to provide that.

1.1 Objective and Contributions

The goal of this paper is to offer a practical framework for studying what people retain from AI-assisted verification. I make three contributions:

1. I define epistemic transfer and show how it differs from immediate correction, trust, reliance, and human–AI team performance.

2. I introduce two complementary quantities: the Epistemic Transfer Effect (ETE), which captures retained independent capability, and Tool-Removal Cost (TRC), which captures current dependence on the tool.

3. I translate these ideas into a concrete evaluation protocol, including conditions, measures, item design, timing, and analysis.

The paper is written as a working paper. The aim is to make the core idea, the design logic, and the proposed protocol as clear and usable as possible.

2 Background and Related Work

2.1 What Current Evaluations Tell Us

Most work on AI-assisted verification answers one of three questions.

First, some studies evaluate the system. Automated fact-checking is often assessed through benchmarks for claim detection, evidence retrieval, and verdict prediction (Guo et al., 2022; Thorne et al., 2018; Schlichtkrull et al., 2023). These studies tell us whether the model performs well in isolation. They do not tell us whether the output improves human judgment, or whether it helps users learn anything.

Second, some studies evaluate the treated claim. Correction and debunking studies ask whether exposure to a correction improves belief accuracy for the claim being corrected (Walter et al., 2020). Studies of labels, provenance signals, and warning cues do something similar: they

measure judgment quality while the cue is present. This is useful, but it still does not tell us how people handle new claims later on.

Third, some studies evaluate the human–AI team. This work looks at team accuracy, trust, reliance, and complementary performance while assistance is available (Lee and See, 2004; Bansal et al., 2021; Vaccaro et al., 2024). Again, this is an important part of the picture, but it mainly tells us what the tool contributes at the point of use.

So the gap is simple. Current evaluations usually tell us whether the system works, whether a cue works, or whether the assisted decision improves. They rarely tell us whether the user becomes better at judging future claims without the system.

2.2 Why Retained Capability Matters

People do not approach information in the same way when tools are available. Cognitive offloading is often useful: external resources can save time and free up mental effort (Risko and Gilbert, 2016). At the same time, access to external support can change how much people search, remember, or reason for themselves (Fisher et al., 2015; Sparrow et al., 2011). That is not necessarily bad. But it means we should not assume that strong assisted performance automatically implies stronger independent performance later.

Automation research makes a similar point. Automation can improve safety and efficiency, while also reducing opportunities to maintain human skill (Bainbridge, 1983). Studies in aviation, for example, show that infrequently practiced flying skills—especially cognitive ones such as navigation and procedure recall—degrade in the automated cockpit (Casner et al., 2014). Meta-analytic work further shows that the effects of automation depend on how much of the task is automated and at what stage (Onnasch et al., 2014). For AI verification tools, the analogous question is not whether the tool helps now, but what kind of user it helps create over time.

2.3 Learning, Retention, and Transfer

Learning research also helps clarify the issue. Doing well during practice is not the same as learning something that lasts (Soderstrom and Bjork, 2015). Easier practice can improve short-term performance without producing the strongest long-term retention. By contrast, retrieval practice and other desirable difficulties may feel harder in the moment while improving later performance (Roediger III and Karpicke, 2006; Bjork and Bjork, 2011).

Transfer adds another requirement. It is not enough that users remember the exact claim they saw before. What matters is whether they can apply a useful strategy to new claims. Transfer can be near or far depending on how much the later task resembles the earlier one (Barnett and Ceci, 2002; Salomon and Perkins, 1989). In verification settings, this may mean using lateral reading, checking multiple sources, or paying attention to source expertise (Wineburg and McGrew, 2019; Kozyreva et al., 2020). Notably, some misinformation interventions are explicitly designed with transfer in mind: inoculation-style "prebunking" games aim to build generalizable resistance to manipulation techniques rather than to correct individual claims (Roozenbeek and van der Linden, 2019). Most AI verification tools are not designed with transfer in mind at all. Whether they nevertheless support transferable strategies—or bypass them—is an empirical question.

2.4 Emerging Evidence from Generative and Clinical AI

Recent studies make the issue more concrete. In high-school mathematics, unrestricted generative AI support improved performance during practice but reduced performance after the tool

was removed; a more scaffolded tutor reduced this problem (Bastani et al., 2025). In misinformation settings, imperfect LLM fact-checking guidance can reduce headline discernment (DeVerna et al., 2024). A month-long dialogue intervention reduced false beliefs about discussed claims, but produced no lasting discernment gains for unseen claims—unassisted accuracy was in fact lower at the end of the study than at baseline (Rani et al., 2026). Other work finds lower persistence and weaker independent performance once AI assistance is withdrawn (Liu et al., 2026). The concern extends beyond generative systems: in the first real-world clinical evidence of this kind, endoscopists who routinely used AI polyp detection showed a significant decline in unassisted adenoma detection, from 28.4% to 22.4%, when performing colonoscopies without the tool (Budzyń et al., 2025). Programming studies also suggest that outcomes depend on whether people remain cognitively engaged or simply hand off the work (Shen and Tamkin, 2026). Related work on incidental learning and self-reported critical thinking points in the same direction (Gajos and Mamykina, 2022; Lee et al., 2025).

The evidence is still mixed. Different studies use different tasks, delays, outcomes, and comparison groups. The point is not that AI necessarily helps or harms learning. The point is that we need a clearer construct and a more standard way to study it.

3 Epistemic Transfer

3.1 Definition

I define *epistemic transfer* as the effect of prior interaction with an AI verification system on a person's later unassisted performance when evaluating new claims.

Three features matter.

First, the later outcome must be measured without the target system or a functionally equivalent AI aid. Otherwise we are still measuring assisted performance.

Second, the later test must use novel claims. Otherwise we may just be measuring memory for a claim or correction seen earlier.

Third, the later test should happen after a retention interval. Otherwise we risk confusing lasting change with short-lived activation.

Epistemic transfer is also comparison-based. A system can help more than no practice, while still helping less than active verification practice. Both comparisons matter. The first tells us whether using the system was better than doing nothing. The second tells us whether the system built or displaced capability relative to the practice it replaced.

This outcome is different from several nearby ideas. A correction effect concerns whether belief accuracy improves for a treated claim (Walter et al., 2020). Trust and reliance concern how people respond to the system while it is present (Lee and See, 2004). Human–AI team performance concerns the quality of the joint decision. None of these tells us, by itself, how well the user performs later on a new claim without the tool.

3.2 Research Questions

The framework is organized around four practical questions:

RQ1: How does prior AI-assisted verification affect delayed unassisted accuracy, calibration, and verification behavior on new claims, compared with active practice and with no practice?

RQ2: How large is the immediate tool-removal cost for different kinds of AI support?

Table 1: Core conditions in the proposed evaluation protocol.

Condition	What the user experiences	Why it is included
Answer-first AI	The system gives a verdict and explanation before the user has to judge independently.	Captures fluent, low-friction assistance that may reduce verification effort.
Evidence-first AI	The system presents sources, uncertainty, and structured prompts before any final verdict.	Tests whether support that keeps the user engaged leads to better transfer.
Active practice	Participants verify comparable claims without AI, using ordinary search and source access.	Shows what users retain from the verification practice that AI may replace.
No-practice control	Participants complete an unrelated matched-duration activity.	Shows whether system use helps more than no practice at all.

RQ3: Which design features—especially answer-first versus evidence-first support—increase or reduce epistemic transfer?

RQ4: How do these effects vary by transfer distance, baseline skill, domain knowledge, and language or media context?

A single study does not need to answer all four questions. But it should keep the core contrast between assisted conditions, active practice, and no practice.

4 Proposed Evaluation Methodology

4.1 Overview

The basic design is a randomized mixed design with four between-participant conditions and a within-participant removal probe. Table 1 shows the core conditions. Figure 1 shows the full study flow. The exact interface and domain can vary, but the basic logic should remain the same: compare different forms of assistance with active practice and with no practice, then test what users can do later on their own.

Extra variants can be added, such as cognitive forcing or different uncertainty displays. But the active-practice control should not be removed. Without it, we cannot tell whether the system builds capability or displaces useful practice.

4.2 Participants and Sampling

Participants should match the intended users of the system. A public-facing tool should be tested with a broad adult sample, not only with students or technically trained users. A professional tool should be tested with the relevant professionals. Multi-country or multilingual studies are especially useful when model quality, source availability, or information environments differ across contexts.

Sample size should be based on the smallest effect that matters in the delayed test. Because responses are nested within participants and items, simulation-based power analysis is preferable

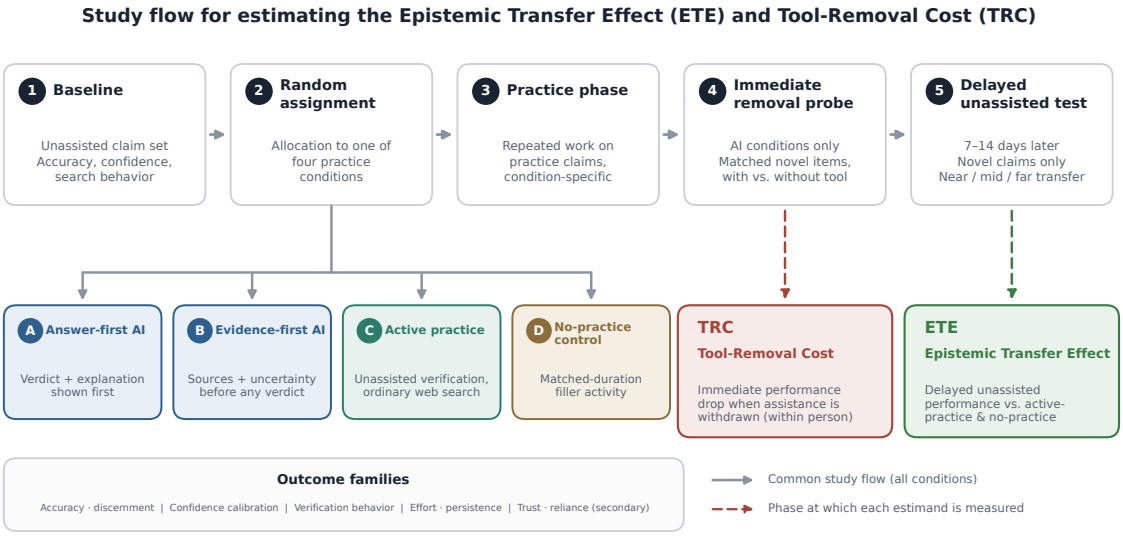

Figure 1: Experimental procedure for estimating epistemic transfer and tool-removal cost. Numbered cards show the common study flow across the five phases; solid arrows denote transitions completed by all participants. The four practice conditions branch from random assignment. Dashed arrows mark the phase at which each estimand is measured: the Tool-Removal Cost (TRC) at the immediate within-person removal probe, and the Epistemic Transfer Effect (ETE) at the delayed unassisted test.

to a simple independent-samples calculation. Eligibility rules, exclusions, attrition handling, and attention checks should be preregistered.

4.3 Claims, Evidence, and Transfer Distance

The claim pool should include true and false claims and, when relevant, mixed or uncertain ones. Claims should come from domains that users plausibly encounter, such as health, politics, climate, or provenance. Each claim should have a documented reference judgment grounded in high-quality evidence.

Claims should be split into baseline, practice, removal-probe, and delayed-test sets. Sets should be matched as well as possible on domain, truth status, familiarity, reading difficulty, and estimated verification difficulty. Pilot work should check for ceiling or floor effects and ensure that the evidence is adequate.

Novelty should be treated carefully. Following transfer research (Barnett and Ceci, 2002), delayed-test claims can be grouped into near, intermediate, and far transfer. Near-transfer claims resemble the practice claims in domain and evidence structure. Intermediate-transfer claims stay in the same domain but require a different kind of evidence. Far-transfer claims require the same verification strategy in a different domain. Textual and semantic overlap checks, together with human review, can help prevent paraphrases of practiced claims from slipping into the delayed test.

4.4 Procedure

The protocol has five phases.

Phase 1: Baseline. Participants judge a balanced set of claims without AI assistance. For each claim, they provide a veracity judgment and a confidence rating. If ordinary web search is part of the task, the study can provide a controlled browser or log search behavior. Baseline measures can also include domain familiarity, prior AI use, verification literacy, and demographics.

Phase 2: Practice. Participants are assigned to one of the four conditions. The practice phase should include enough trials—or enough repeated sessions—to make learning or displacement plausible.

Phase 3: Immediate removal probe. Participants in the AI conditions complete matched novel items both with and without the assigned tool. Tool availability and item order should be randomized or counterbalanced within participant. This phase estimates TRC. One design detail deserves attention: the probe itself involves unassisted work on novel items and can therefore act as an additional learning event. To keep exposure comparable across conditions, the active-practice and no-practice groups should complete a matched unassisted probe block of equal length. Otherwise, differences at the delayed test could partly reflect the probe rather than the practice phase.

Phase 4: Delayed unassisted test. After a pre-specified delay, for example 7–14 days, all participants judge new claims without AI assistance. A second follow-up around four weeks later can be added to study decay.

Phase 5: Debriefing. After the final outcome measurement, participants should receive reliable corrections and source information, especially if the study used false health or political claims.

4.5 Outcome Measures

No single outcome is enough here. Verification quality has several parts.

Accuracy and discernment. The main outcome can be binary accuracy, an ordinal veracity judgment, or a continuous credibility rating. For balanced true and false claims, discernment can be defined as the difference in mean credibility assigned to true versus false claims.

Confidence calibration. Participants should report how confident they are in each judgment. A system may leave accuracy unchanged while increasing overconfidence, which would be an important negative outcome.

Verification behavior. Behavioral measures can include whether participants search, how many sources they open, how diverse those sources are, how much time they spend on evidence, and whether they use lateral reading or inspect evidence that conflicts with their first impression (Wineburg and McGrew, 2019).

Effort and persistence. Time on task, number of search actions, abandonment, and willingness to continue after difficulty can help show whether the tool changes engagement or persistence (Liu et al., 2026; Lee et al., 2025).

Secondary outcomes. Trust, usefulness, satisfaction, reliance, and intention to reuse are still worth measuring, but they should not be confused with epistemic transfer.

4.6 Analysis Plan

The main analysis should use mixed-effects models, because claims are nested within participants and items. For a binary accuracy outcome, one suitable specification is

$$\text{logit}[\Pr(Y_{ij} = 1)] \;=\; \beta_0 + \beta_1 C_i + \beta_2 D_j + \beta_3 \left(C_i \times D_j\right) + \gamma^\top X_i + u_i + v_j, \tag{1}$$

where Y_{ij} is the accuracy of participant i on claim j, C_i is condition, D_j is transfer distance, X_i contains preregistered baseline covariates, and u_i and v_j are participant and item random intercepts.

The two main ETE contrasts are straightforward: each AI condition versus active practice, and each AI condition versus no practice. Effects should be reported as marginal effects or predicted probabilities with confidence intervals, not only as raw coefficients. If the goal is to argue that transfer is practically null, equivalence tests should be used rather than treating a non-significant difference as proof of no effect (Lakens, 2017).

TRC is estimated within the AI conditions using the immediate with-tool versus without-tool comparison. That model should include tool availability, design condition, their interaction, item set, and order. A joint analysis can then ask whether larger TRC is associated with lower delayed ETE, while recognizing that such associations may not themselves be causal.

4.7 Feasibility

A fair question is who can realistically run this protocol. The design is heavier than a single-session experiment, but it is well within the reach of standard online-panel research. As an illustration, detecting a delayed between-condition difference of three to five percentage points in accuracy—with 20–30 delayed-test items, participant and item random effects, and conventional error rates—will typically require several hundred participants per condition in simulation-based power analyses; a four-condition study should therefore plan for roughly 1,200–2,000 participants in total, depending on item counts, intraclass correlations, and the smallest effect of interest. Two-wave designs with a 7–14 day delay routinely retain a large majority of panel participants (Kothe and Ling, 2019), so oversampling at recruitment by around a quarter is a reasonable default. These figures are illustrative rather than prescriptive: the point is that the protocol is a normal-sized preregistered online experiment with one extra wave, not a bespoke laboratory program.

5 Two Complementary Estimands

5.1 Epistemic Transfer Effect

For AI condition c, comparator k, transfer distance d, and assessment regime b, I define

$$\text{ETE}(c, k, d, b) \;=\; \mathbb{E}[Y_{\text{delay}} \mid c, d, b, X_0] \;-\; \mathbb{E}[Y_{\text{delay}} \mid k, d, b, X_0]\,, \tag{2}$$

where Y_{delay} is delayed unassisted performance on novel claims and X_0 contains preregistered baseline covariates. In plain language, ETE asks: after the delay, how much better or worse do users perform without the tool compared with users in another condition?

ETE should be reported separately against active practice and no practice. A positive effect relative to no practice means that prior system use helped more than doing nothing. A negative effect relative to active practice means that the system produced less retained capability than the verification activity it displaced.

5.2 Tool-Removal Cost

For AI condition c, I define

$$\mathrm{TRC}(c) \;=\; \mathbb{E}[Y_{\text{probe}} \mid \text{tool available}, c] \;-\; \mathbb{E}[Y_{\text{probe}} \mid \text{tool removed}, c]\,. \tag{3}$$

TRC captures the immediate advantage of having the tool available on matched novel items. In plain language, it asks: how much worse do users do when the system is suddenly not there?

TRC is related to, but distinct from, complementarity in the human–AI teaming literature (Bansal et al., 2021; Vaccaro et al., 2024). Complementarity typically asks whether the human–AI team outperforms the human or the AI alone, usually in a single session and across participants. TRC instead isolates the tool's marginal contribution within the same person, on matched novel items, after a practice phase—and it is designed to be read jointly with ETE. The same numerical tool advantage means something different when it coexists with strong retained capability than when it does not. That joint reading is the point of the diagnostic space below.

TRC is not, by itself, a measure of harm. A professional tool may create a large immediate advantage while still helping users build useful skill. That is why ETE and TRC are most informative when interpreted together.

6 Diagnostic Space

Figure 2 crosses ETE against active practice with TRC. This yields four descriptive profiles.

Capability building. Users retain improved performance and are not strongly dependent on the tool for current performance.

Capability plus tool advantage. Users learn, but the system still gives them an extra boost when it is available. This may be a realistic target for professional decision support.

Verification on loan. Users perform better while the system is present, but that benefit is not retained relative to active practice. The performance is effectively rented from the tool.

Epistemically inert or de-skilling. The system provides little immediate advantage and does not improve retained capability; a negative ETE relative to active practice is consistent with displacement of useful practice.

This framework helps avoid two common mistakes: treating strong assisted performance as evidence of learning, and treating every tool advantage as unhealthy dependence.

7 Implications for Research, Design, and Policy

7.1 Research

The framework gives researchers a common language for comparing studies that currently use different concepts and endpoints. At a minimum, studies should report the comparator, delay, access regime, item novelty, transfer distance, and outcome family alongside any transfer estimate. Without these details, two studies may appear to disagree while in fact estimating different things.

The framework also makes heterogeneity a first-class question. Effects may differ by baseline skill, age, domain expertise, language, or media environment. Some users may benefit most from immediate answers, while also losing the most opportunity to practice verification.

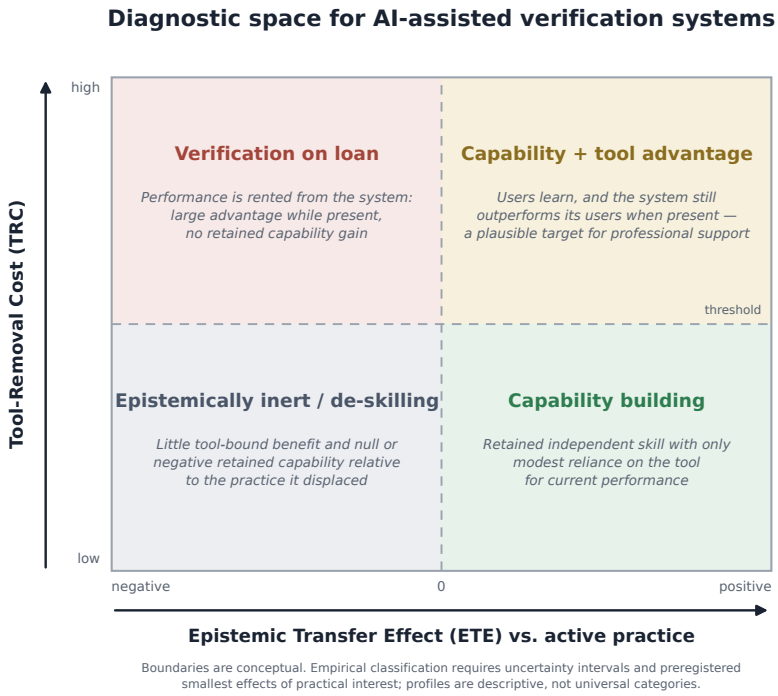

Figure 2: Diagnostic space defined by the Epistemic Transfer Effect against active practice and Tool-Removal Cost. Boundaries are conceptual. Empirical classification requires uncertainty intervals and preregistered smallest effects of practical interest.

7.2 System Design

The design question is not just whether to add explanations. It is whether the interface keeps users cognitively involved in ways that may later matter. Answer-first interfaces, evidence-first interfaces, cognitive forcing, uncertainty displays, and required user participation can all be compared using the same ETE and TRC outcomes. Existing work suggests that engagement and workflow cost shape overreliance and incidental learning (Buçinca et al., 2021; Gajos and Mamykina, 2022; Vasconcelos et al., 2023).

This also makes trade-offs visible. A more demanding interface may reduce speed or satisfaction while improving retained capability. A very fluent interface may maximize short-term performance while producing little transfer. These are empirical trade-offs, and they should be reported as such.

7.3 Policy and Procurement

The EU AI Act includes provisions related to AI literacy, human oversight, and transparency (European Parliament and Council of the European Union, 2024). This paper does not claim that epistemic transfer is itself a legal requirement. But it does offer a useful evaluation idea for settings where organizations claim that an AI system supports human competence or informed judgment.

A simple procurement checklist follows from the framework. Decision makers can ask: Does the system improve immediate decisions? How much does performance drop when the system

is unavailable? Does repeated use improve or weaken later independent performance? The colonoscopy case above illustrates why these questions matter: a deployed clinical AI system was associated with measurably worse unassisted performance, a pattern such questions are designed to surface before deployment (Budzyń et al., 2025). Different answers may call for different responses, including interface redesign, more realistic claims about the tool, fallback procedures, or periods of unassisted practice.

8 Boundary Conditions and Limitations

Epistemic transfer is not equally relevant for every tool. It matters most when use is repeated, the task contains learnable strategies, users are likely to face similar situations without equivalent assistance, and errors have meaningful consequences. A calculator can be valuable without teaching arithmetic. Likewise, a specialized professional tool may be useful even if users cannot match its full performance independently.

The proposed measures also have limits. Novel claims may still resemble practiced claims in ways that are hard to detect. Transfer distance is domain-specific. Process measures can change when the study constrains normal search behavior. The removal probe, even when matched across conditions as recommended above, adds unassisted retrieval practice for everyone and may therefore slightly compress condition differences at the delayed test; designs that omit the probe estimate ETE without this influence but give up TRC. The diagnostic space should therefore not be reduced to a single universal certification score.

Tool-removal probes also require judgment. Assistance should not be withheld during real, high-stakes decisions if participants would ordinarily be entitled to use it. In such cases, removal is better studied through simulations, retrospective tasks, or other low-risk exercises. Finally, the framework itself is still a proposal. It needs to be tested, refined, and compared across domains.

9 Conclusion

AI-assisted verification is usually evaluated by what happens while the tool is present. This paper argues that, when independent judgment matters, that is not enough. We also need to know what users can still do later on their own.

To study this, I introduce the idea of epistemic transfer and propose two complementary quantities: ETE for delayed unassisted performance and TRC for immediate dependence on the tool. I then turn these ideas into a practical evaluation protocol with clear conditions, measures, and analytic choices.

The main claim is modest but important. AI assistance does not have to teach in every setting. But we should stop assuming that strong point-of-use performance tells us what users learn. Sometimes the tool may build capability. Sometimes it may provide only a temporary boost. And sometimes it may do both. The only way to know is to measure it.

Declarations

Funding. This work was supported in part by the Research Council of Norway through SFI MediaFutures, Research Centre for Responsible Media Technology and Innovation (grant no. 309339).

Competing interests. The author declares no competing interests.

AI-assisted drafting. Generative AI tools were used to support drafting and editing.

References

Lisanne Bainbridge. Ironies of automation. *Automatica*, 19(6):775–779, 1983. https://doi.org/10.1016/0005-1098(83)90046-8.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021. Article 81. https://doi.org/10.1145/3411764.3445717.

Susan M. Barnett and Stephen J. Ceci. When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, 128(4):612–637, 2002. https://doi.org/10.1037/0033-2909.128.4.612.

Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakçı, and Rei Mariman. Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26):e2422633122, 2025. https://doi.org/10.1073/pnas.2422633122.

Elizabeth L. Bjork and Robert A. Bjork. Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In Morton Ann Gernsbacher, Richard W. Pew, Leaetta M. Hough, and James R. Pomerantz, editors, *Psychology and the Real World*, pages 56–64. Worth, 2011.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof 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(CSCW1):1–21, 2021. Article 188. https://doi.org/10.1145/3449287.

Krzysztof Budzyń, Marcin Romańczyk, Diana Kitala, Paweł Kołodziej, Marek Bugajski, Hans Olov Adami, Johannes Blom, Marek Buszkiewicz, Natalie Halvorsen, Cesare Hassan, Tomasz Romańczyk, Øyvind Holme, Krzysztof Jarus, Shona Fielding, Melina Kunar, Maria Pellise, Nastazja Pilonis, Michał Filip Kamiński, Mette Kalager, Michael Bretthauer, and Yuichi Mori. Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. *The Lancet Gastroenterology & Hepatology*, 10(10):896–903, 2025. https://doi.org/10.1016/S2468-1253(25)00133-5.

Stephen M. Casner, Richard W. Geven, Matthias P. Recker, and Jonathan W. Schooler. The retention of manual flying skills in the automated cockpit. *Human Factors*, 56(8):1506–1516, 2014. https://doi.org/10.1177/0018720814535628.

Matthew R. DeVerna, Harry Yaojun Yan, Kai-Cheng Yang, and Filippo Menczer. Fact-checking information from large language models can decrease headline discernment. *Proceedings of the National Academy of Sciences*, 121(50):e2322823121, 2024. https://doi.org/10.1073/pnas.2322823121.

European Parliament and Council of the European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, 2024. https://data.europa.eu/eli/reg/2024/1689/oj.

Matthew Fisher, Mariel K. Goddu, and Frank C. Keil. Searching for explanations: How the Internet inflates estimates of internal knowledge. *Journal of Experimental Psychology: General*, 144(3):674–687, 2015. https://doi.org/10.1037/xge0000070.

Krzysztof Z. Gajos and Lena Mamykina. Do people engage cognitively with AI? Impact of AI assistance on incidental learning. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, pages 794–806, 2022. https://doi.org/10.1145/3490099.3511138.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 2022. https://doi.org/10.1162/tacl_a_00454.

Emily J. Kothe and Mathew Ling. Retention of participants recruited to a one-year longitudinal study via Prolific. PsyArXiv, 2019. https://doi.org/10.31234/osf.io/5yv2u.

Anastasia Kozyreva, Stephan Lewandowsky, and Ralph Hertwig. Citizens versus the Internet: Confronting digital challenges with cognitive tools. *Psychological Science in the Public Interest*, 21(3):103–156, 2020. https://doi.org/10.1177/1529100620946707.

Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017. https://doi.org/10.1177/1948550617697177.

Hao-Ping Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–22, 2025. Article 1121. https://doi.org/10.1145/3706598.3713778.

John D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004. https://doi.org/10.1518/hfes.46.1.50_30392.

Grace Liu, Brian Christian, Tsvetomira Dumbalska, Michiel A. Bakker, and Rachit Dubey. AI assistance reduces persistence and hurts independent performance. arXiv:2604.04721, 2026. https://doi.org/10.48550/arXiv.2604.04721.

Linda Onnasch, Christopher D. Wickens, Huiyang Li, and Dietrich Manzey. Human performance consequences of stages and levels of automation: An integrated meta-analysis. *Human Factors*, 56(3):476–488, 2014. https://doi.org/10.1177/0018720813501549.

Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2):230–253, 1997. https://doi.org/10.1518/001872097778543886.

Anku Rani, Valdemar Danry, Paul Pu Liang, Andrew B. Lippman, and Pattie 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*, pages 1–26, 2026. Article 792. https://doi.org/10.1145/3772318.3790656.

Evan F. Risko and Sam J. Gilbert. Cognitive offloading. *Trends in Cognitive Sciences*, 20(9): 676–688, 2016. https://doi.org/10.1016/j.tics.2016.07.002.

Henry L. Roediger III and Jeffrey D. Karpicke. Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3):249–255, 2006. https://doi.org/10.1111/j.1467-9280.2006.01693.x.

Jon Roozenbeek and Sander van der Linden. Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, 5, 2019. Article 65. https://doi.org/10.1057/s41599-019-0279-9.

Gavriel Salomon and David N. Perkins. Rocky roads to transfer: Rethinking mechanisms of a neglected phenomenon. *Educational Psychologist*, 24(2):113–142, 1989. https://doi.org/10.1207/s15326985ep2402_1.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. AVeriTeC: A dataset for real-world claim verification with evidence from the web. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*, 2023. https://doi.org/10.48550/arXiv.2305.13117.

Judy Hanwen Shen and Alex Tamkin. How AI impacts skill formation. arXiv:2601.20245, 2026. https://doi.org/10.48550/arXiv.2601.20245.

Nicholas C. Soderstrom and Robert A. Bjork. Learning versus performance: An integrative review. *Perspectives on Psychological Science*, 10(2):176–199, 2015. https://doi.org/10.1177/1745691615569000.

Betsy Sparrow, Jenny Liu, and Daniel M. Wegner. Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, 333(6043):776–778, 2011. https://doi.org/10.1126/science.1207745.

Benjamin C. Storm and Sean M. Stone. Saving-enhanced memory: The benefits of saving on the learning and remembering of new information. *Psychological Science*, 26(2):182–188, 2015. https://doi.org/10.1177/0956797614559285.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: A large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 809–819, 2018. https://doi.org/10.18653/v1/N18-1074.

Michelle Vaccaro, Abdullah Almaatouq, and Thomas W. Malone. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8:2293–2303, 2024. https://doi.org/10.1038/s41562-024-02024-1.

Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7 (CSCW1):1–38, 2023. Article 129. https://doi.org/10.1145/3579605.

Nathan Walter, Jonathan Cohen, R. Lance Holbert, and Yasmin Morag. Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, 37(3):350–375, 2020. https://doi.org/10.1080/10584609.2019.1668894.

Sam Wineburg and Sarah McGrew. Lateral reading and the nature of expertise: Reading less and learning more when evaluating digital information. *Teachers College Record*, 121(11):1–40, 2019. https://doi.org/10.1177/016146811912101102.