

Verification on Loan

When AI Checks Facts So We Don't Have To

Christoph Trattner

University of Bergen, Norway
christoph.trattner@uib.no

ABSTRACT

AI-assisted verification, in which a system checks whether a claim or an image can be trusted, can improve users' immediate judgments. Whether users keep the competence to notice when such a system is wrong is rarely measured. This article argues that the duty to find out falls on those who build and deploy verification systems. It proposes three obligations: measure unaided performance after sustained use, disclose a dependency profile, and offer users meaningful participation.

INTRODUCTION

In this column in early 2024 [5], my co-authors and I argued that professional norms could help translate abstract AI principles into practice. Mature professions such as journalism, we suggested, have spent decades turning values into working rules.

Since then, a different ethical problem has become harder to ignore. The question is no longer only whether an AI system behaves responsibly. It is also what happens over time to the people who rely on it when it performs its task well.

Consider AI-assisted verification, in which a system checks, for anyone who asks, whether a claim, an image, or a news story can be trusted. Systems now assess claims, retrieve evidence, flag synthetic media, and attach provenance information – for example through the C2PA standard, which records media origin and editing history. Others go further, judging a claim true, false, misleading, or uncertain.

This is largely welcome. Misinformation causes real harm, and ordinary users cannot investigate every claim they encounter. But verification cannot be delegated without residue. It involves recognizing when scrutiny is needed, finding evidence, understanding uncertainty, and calibrating confidence.

When an AI system performs those operations for us, it lends us its competence: with it, we judge better than we could alone. The ethical question is what remains when the loan is withdrawn, because the system is unavailable, uncertain, or wrong, and who is responsible for finding out.

WHAT REMAINS AFTER ASSISTANCE?

The first half of that question is not new. In 1991, Salomon and colleagues distinguished the effects of working with an intelligent technology from the effects of it: what a person achieves in partnership with the machine, and what that person can do afterwards without it [9]. Cognitive-offloading research has since shown that external tools can improve immediate performance while changing what people themselves remember [8]. Vallor has argued that moral skills, such as those of caring for others, are also learned through practice and can fade in the same way [10].

AI verification can therefore produce different outcomes even when the user’s judgment of the claim at hand is correct. An explanation that exposes sources, uncertainty, and limits may invite the user to participate in the judgment. A verdict shown without its evidence may remove the need to inspect anything. A provenance record, likewise, can be shown as an inspectable history or as a green badge that users learn to accept unread. Both kinds of design can help the user answer correctly today. Only the first may leave the user better equipped tomorrow. The interface decision is an ethics decision, because it affects what users can later do for themselves.

Evidence on which of the two we are getting is thin. In a month-long CHI 2026 study, participants judged whether news items were real or fake: alone, then with an AI assistant, then alone again on new items [7]. The assistant made them more accurate, but alone they did not improve. Tested directly after an assisted round, their accuracy in week four was 15.3 percentage points lower than at the outset. Tested at the start of a session, which the authors consider the better measure of learning, it was 6.8 points lower, which was not statistically significant; they note possible fatigue and recovery between sessions. The larger drop was largely confined to fake items, which may reflect a growing readiness to accept items as real rather than lost discernment. The study thus shows no lasting gain in unaided judgment, and whether such assistance erodes it over months of ordinary use is unknown.

There is a second complication: the verification system can be wrong. In a preregistered experiment, fact-checks generated by a large language model did not significantly improve overall headline discernment as human fact-checks did. When the model mislabeled or expressed uncertainty about particular headlines, it could move participants in the wrong direction [4].

We should resist overreaction to such findings. People have always relied on others and institutions for knowledge they cannot establish alone, and automation has always retired skills. Few of us can wash clothes by hand as our grandparents could, and nobody is wronged by that. Verification differs in one respect: I can see that a washing machine has failed without knowing how to wash, but I cannot see that a verification system has failed without some of the competence it replaces. Epistemic autonomy therefore means retaining enough competence to know when scrutiny is needed, which sources are appropriate, and when reliance is justified.

WHY THIS IS AN ETHICS PROBLEM

The long-term effect of a system on its user is also an ethical property of the system.

First, there is autonomy. Major AI ethics frameworks put human agency and oversight near the center of responsible design [6]. The ACM Code likewise requires computing professionals to contribute to human well-being and avoid harm [1]. Yet autonomy is more than having the final say. A person remains formally free to decide while substantively

dependent on a system supplying evidence, interpretation, and conclusion. If repeated use weakens the capacity to assess those outputs, “human in the loop” becomes a procedural description rather than meaningful oversight.

Second, there is responsibility. Developers often describe AI output as assistance while leaving users responsible for the decisions that follow. That allocation becomes hard to defend when a product is designed so that users routinely accept what it says. Those who engineer interfaces that compress evidence into authoritative conclusions and cultivate dependence through convenience cannot then treat each user as an independent verifier whenever the system fails.

The result would be a responsibility gap created not by the absence of a human decision, but by the gradual weakening of the human capacity presupposed by that decision.

A DUTY TO PRESERVE CAPABILITY

The ethical response is not to reject AI-assisted verification, nor to restate an observation that is three decades old: performing well with a tool differs from being able to perform without it. The practical question is who must act on it. My claim is that a capability-preservation duty falls on those who build and deploy verification systems: they design the interface, set the conditions of reliance, and hold the data needed to evaluate its effects at scale.

The press offers a precedent. Nobody supposes that newspapers owe readers a reporter’s skills, yet journalism accepted duties that keep readers able to judge it: naming sources, correcting errors publicly, and separating news from comment. The duty proposed here is of that kind. It entails three obligations, and the first two ask nothing of users.

Measure the user after the tool. Verification systems are typically evaluated by classification accuracy, retrieval quality, user trust, or immediate belief correction. None shows whether a system has built competence, created dependence, or done neither. Evaluations should compare unaided performance before and after sustained use, on unfamiliar claims, after a delay, and against a group that worked without the tool. They should also test whether users still reject the system’s advice when it is wrong. The CHI study had no comparison group without AI, and its result depended on when the test was taken.

Disclose the dependency profile. Users, and the institutions that buy on their behalf, deserve to know what a product has been tested to do to them over time. Dependency here means relying on output that users cannot check, because the skill has faded or was never built. A dependency profile would state what was tested, over what period, after what delay, and against which comparison. It would report the change in unaided performance and in the detection of erroneous AI advice, each with its margin of error. An entry of “not tested” would be honest, but it would not discharge the duty to measure. Where testing shows that unaided judgment deteriorates, the design should be changed and tested again. Labels for cognitive effects may sound fanciful, but so once did the privacy labels that app stores now require.

Offer meaningful participation. Systems can offer graduated assistance, for example by showing the evidence first and the verdict only on request. Designs that withhold or delay the AI’s answer, called cognitive forcing functions, reduce overreliance on AI advice; whether they preserve unaided judgment is for the first obligation to test. Yet this is where my proposal becomes paternalistic, and it is better to say so. Many users want the verdict, not the evidence, and will click straight to it, as they click past cookie

banners. In an experiment with such designs, participants liked least the variants that reduced overreliance most, and those who enjoy effortful thinking benefited more [2]. A general mandate would therefore override what users want while benefiting them unequally. I defend something narrower: participation as a default that users can switch off, and as a requirement only where an institution is responsible for its members' competence, as newsrooms and schools are. Elsewhere the less paternalistic route runs through the first two obligations, so that people who choose convenience can do so informed.

These obligations run against commercial incentives, which follow what users reward. That is why capability preservation belongs in professional standards such as ACM's, in the conditions public bodies set when buying such systems, and in research funders' requirements, rather than being left to product metrics.

A LESSON FROM AVIATION

When airline pilots accustomed to automated cockpits were tested in a simulator, their manual control and instrument scanning were largely intact. The problems lay in the cognitive tasks of manual flight, such as tracking position without a map display and diagnosing instrument failures [3]. These are problems of judgment rather than of hand and eye, which makes them the closer analogue to verification. The authors recommended more frequent practice of these skills, including in recurrent training, and more active monitoring of the automation [3].

Unlike pilots, most users of verification systems receive no recurrent training, and no institution is responsible for ensuring that they retain the ability to evaluate evidence once automated systems become the normal interface to information. Design must take the place of training, and evaluation must show whether it does.

We are increasingly lending users a competence they may never build themselves. A responsible loan should not be structured so that the borrower remains permanently in debt. The goal is not to make every user an expert fact-checker. It is to ensure that convenience does not quietly eliminate the minimum competence required for informed reliance.

The next generation of verification systems should therefore be judged by two questions: Did the system help the user reach a better conclusion? And, after repeated use, did the user remain capable of judging when the system was absent, uncertain, or wrong? Those who build and deploy these systems should be expected to provide the evidence for the second.

REFERENCES

- [1] Association for Computing Machinery. 2018. ACM Code of Ethics and Professional Conduct. <https://www.acm.org/code-of-ethics>
- [2] Z. Bućinca, M. B. Malaya, and K. 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, 1–21. <https://doi.org/10.1145/3449287>
- [3] S. M. Casner, R. W. Geven, M. P. Recker, and J. W. Schooler. 2014. The retention of manual flying skills in the automated cockpit. *Human Factors* 56, 8, 1506–1516. <https://doi.org/10.1177/0018720814535628>
- [4] M. R. DeVerna, H. Y. Yan, K.-C. Yang, and F. Menczer. 2024. Fact-checking information from large language models can decrease headline discernment. *Proceedings of the National Academy of Sciences* 121, 50, e2322823121. <https://doi.org/10.1073/pnas.2322823121>

- [5] N. Diakopoulos, C. Trattner, D. Jannach, I. Costera Meijer, and E. Motta. 2024. Leveraging professional ethics for responsible AI. *Communications of the ACM* 67, 2 (Feb. 2024), 19–21. <https://doi.org/10.1145/3625252>
- [6] High-Level Expert Group on Artificial Intelligence. 2019. Ethics Guidelines for Trustworthy AI. European Commission, Brussels.
- [7] A. Rani, V. Danry, P. P. Liang, A. B. Lippman, and P. 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 (CHI '26)*, Article 792, 1–26. <https://doi.org/10.1145/3772318.3790656>
- [8] E. F. Risko and S. J. Gilbert. 2016. Cognitive offloading. *Trends in Cognitive Sciences* 20, 9, 676–688. <https://doi.org/10.1016/j.tics.2016.07.002>
- [9] G. Salomon, D. N. Perkins, and T. Globerson. 1991. Partners in cognition: Extending human intelligence with intelligent technologies. *Educational Researcher* 20, 3, 2–9. <https://doi.org/10.3102/0013189X020003002>
- [10] S. Vallor. 2015. Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character. *Philosophy & Technology* 28, 1, 107–124. <https://doi.org/10.1007/s13347-014-0156-9>

ABOUT THE AUTHOR

Christoph Trattner is a full professor of information science and media studies at the University of Bergen, Norway, and founder and director of SFI MediaFutures, a Research Centre for Responsible Media Technology and Innovation.

AI-use disclosure: The author used Claude (Anthropic) to assist in performing the research for and the writing and revision of this article, including refining its language, and ChatGPT (OpenAI) to assist with revision and editing.