

# "AI Makes Mistakes" Is Not an Argument

The existence of AI error is a warning, not a complete argument. This article asks newsrooms to compare whole workflows on actual tasks and evaluate the severity, detectability, overlap and propagation of their errors.

2026.10.20 18:46 DENNIS HEDEGREEN ANALYSE V1.0

[HTTPS://HEDEGREENRESEARCH.COM/ARTICLES/AI-MAKES-MISTAKES-IS-NOT-AN-ARGUMENT/](https://hedegreenresearch.com/articles/ai-makes-mistakes-is-not-an-argument/)

---

**The question is not whether AI makes mistakes. The question is which workflow leaves the important mistakes uncaught.**

AI makes mistakes. That part is easy. Models hallucinate, misread context, invent citations, flatten uncertainty and confidently produce answers that sound much better than they deserve. None of that should be waved away.

But “AI makes mistakes” is still not a complete argument against using AI in journalism. It is a warning. To become an argument, it needs a task, a failure mode and a comparison.

The relevant question is not:

Does AI make mistakes?

Of course it does. The relevant question is:

**Compared with what workflow, doing what task, under what controls?**

That sounds less dramatic than “AI is unreliable,” but it is much more useful.

## Stop comparing species

---

A lot of AI debate still falls into a strangely primitive contest:

human vs AI

Which one is better? That is usually the wrong unit. Humans are not one system, and neither is AI. A veteran investigative reporter with three weeks, source access and domain expertise is not equivalent to an exhausted junior reporter working against a deadline. Likewise, a current model with retrieval, tools and source links is not equivalent to a bare chatbot from three years ago.

So the useful comparison is not:

human versus AI

It is:

workflow A versus workflow B

For example:

**Workflow A** A reporter manually searches 20,000 documents.

**Workflow B** AI searches the 20,000 documents for candidate matches, then the reporter opens the original records and verifies every relevant finding.

Those workflows can fail in different ways. The human may miss the document entirely; the AI may surface the wrong one. The question is not which participant is philosophically more trustworthy. The question is which workflow leaves fewer important mistakes uncaught.

**The unit of comparison should be the workflow, not the actor.**

## **Accuracy is not enough**

---

Even that is not enough, because “accuracy” can hide the errors that actually matter. Imagine two workflows. Workflow A makes ten minor transcription mistakes. Workflow B makes two mistakes, but one of them falsely attributes a criminal allegation to the wrong person. Workflow B has the better error count. It is still the workflow I would least like attached to my byline.

**Ten harmless mistakes and one catastrophic mistake do not belong in the same denominator.**

A useful error model therefore needs more than frequency. It needs at least:

- how severe the error is;
- how easy it is to detect;
- whether it propagates;
- whether another part of the workflow can catch it.

A typo corrected before publication is not the same kind of failure as a fabricated quotation sent to a million readers.

Average accuracy can become actively misleading when the rare failure is the one the newsroom cannot afford.

## **Human + AI is not automatically better**

---

There is another comforting slogan that does not survive contact with the evidence:

Human + AI is best.

Sometimes it is. Sometimes it is not.

A 2025 meta-analysis of 146 experiments in healthcare and public-sector decision settings found reliable human augmentation from AI assistance (Hedges'  $g = 0.622$ ), while synergy remained negative overall (Hedges'  $g = -0.380$ ). The authors also found substantial heterogeneity and publication bias favouring

positive augmentation results. Even with that qualification, the distinction matters: AI can make a human better without creating a combined system that beats the better solo performer. [1]

Field experiments show the same reason for caution from another angle.

In a 2026 preregistered field experiment involving 791 Procter & Gamble professionals, with complete data from 776, participants worked on real product-innovation challenges. Individuals using generative AI matched the performance of teams working without it, and AI helped narrow the difference between commercially oriented and technically oriented proposals. [2]

Three randomized field experiments involving 4,867 software developers at Microsoft, Accenture and another large company found roughly 26% more completed tasks with AI coding assistance. [3]

Those are real gains. They are also not proof that AI makes journalism more accurate. Productivity is not accuracy, and finishing more tasks is not the same thing as making fewer factual errors.

The evidence is useful because it proves a narrower point:

**A fallible tool can improve a fallible process.**

It does not tell us, by itself, whether that process is safe enough for a particular journalistic task.

## **Agreement is not verification**

---

One of the nastiest human–AI failure modes is also one of the easiest to mistake for reassurance: agreement.

Suppose a reporter misreads a paragraph. The AI summary makes the same mistake. The reporter asks the AI to check the interpretation, and the AI agrees. Now we have two systems saying the same thing, which feels comforting. Unfortunately, it may still just be one mistake wearing two hats.

Research on human–AI decision-making distinguishes several useful interaction errors.

A **false confirmation** occurs when human and AI agree and both are wrong. A **false conflict** occurs when the human starts out correct, the AI is wrong, and the human switches to the AI's answer. A **true conflict** occurs when the AI is correct, the human is wrong, and the human rejects the AI. [4] [5]

Those categories come from medical decision research, so they are not newsroom evidence. But they give journalism a useful way to think about interaction errors: agreement can be wrong, correct humans can be pulled toward wrong AI advice, and correct AI advice can be rejected.

The important point is not that humans and AI disagree. It is whether the disagreement leads back to evidence.

**Agreement is not independent verification when both systems inherit the same mistake.**

That is why a second opinion only helps if it is meaningfully independent. Two readers of the same bad source do not create a better source. Two systems trained or prompted into the same frame may simply reproduce the same failure with greater confidence.

## Independence should be designed, not assumed

---

This creates an awkward problem for “human in the loop” systems. If the AI speaks first, the human review may already be contaminated by the model's framing. That does not mean every workflow should force a reporter to work blind before consulting AI; that would quickly become ritualistic and annoying. But in high-stakes tasks, order can matter.

Compare:

AI answer -> human approves

with:

human initial judgement -> AI challenge -> original source -> resolution

The second process creates more opportunity for disagreement to become useful. The real check is still the source. The human and the AI are not supposed to vote on reality; they are supposed to help the newsroom find and interpret evidence.

**Independence should be designed, not assumed.**

## Complementarity is a workflow property

---

This is where “human + AI” becomes too vague to be useful. There are many possible architectures:

- AI searches; human verifies.
- AI translates; human checks the quote.
- Human reports; AI searches for contradictions.
- AI drafts; human edits.
- Both make a full decision; the human picks one.

Those are not one thing.

A 2026 PNAS Nexus framework on human–AI teaming puts much more emphasis on task structure, role partitioning, trust calibration and the conditions for genuine complementarity than on generic “teamwork.” [6]

That makes sense. If both systems are asked to solve the same problem in the same way, they may reproduce the same errors. If they do different work, one may compensate for the other's weakness.

For journalism, one of the cleanest examples is document search. The AI can inspect a corpus too large for a human team to read exhaustively. The human can then inspect the original passages, understand context, judge materiality and take responsibility for the claim. That is a complementary architecture.

It is very different from:

AI writes answer -> human checks whether it feels right

which is basically peer review by vibes with extra electricity.

**Workflow architecture determines whether the systems complement or contaminate each other.**

## **Rare failures can dominate the decision**

---

There are tasks where one error is enough: a fabricated quote, the wrong identity, a false allegation, a leaked confidential source, or an invented citation in a legal or scientific story.

For those tasks, a model can have an impressive average score and still be unacceptable without hard verification.

This is why benchmark percentages are not newsroom policy. A tool that is 99.9% correct may be excellent for one workflow and unusable for another. If the remaining 0.1% produces formatting errors, perhaps nobody cares. If it can accuse the wrong person of a crime, we suddenly care quite a lot.

**Average accuracy is useless if the rare failure is the one your newsroom cannot afford.**

The correct control may be:

- mandatory source verification;
- restricted use;
- human-only final judgement;
- or simply not using the model for that task.

The point is not to maximize AI use. It is to minimize uncaught important error.

## **Today's evidence has a half-life**

---

There is another problem with the human–AI literature: the subject keeps changing. A study from 2024 may still tell us something valuable about anchoring, reliance, false confirmation or workflow design. But a concrete performance result from 2024 is not automatically a good description of a 2026 model, a 2026 interface or a 2026 workforce that has had two more years to learn how to use the tools.

That is not a reason to throw old research away. It is a reason to separate mechanisms from current performance.

**Human–AI evidence has a half-life.**

The models, interfaces, context windows and patterns of tool use change. The humans change too.

We are not only testing immature AI. We are testing immature ways of working with AI.

Humanity has, in a sense, managed to light the fire and is currently arguing about which logs burn best. That is useful work. It is still a little early to publish the final handbook on metallurgy.

The smartphone is a useful reminder. In one generation, human habits reorganized around a device that barely existed before. There is no reason to assume today's prompt box, today's job descriptions or today's "human reviews AI answer" workflow is the final human–AI arrangement.

That makes current studies important, and it makes their scope important.

A result should carry:

- the model generation;
- the task;
- the interface;
- the user's expertise;
- the date.

In a fast-moving field, "AI performs X% better" without those details ages badly.

## **We may be testing AI inside inherited jobs**

---

There is a second layer to that immaturity. The field experiments discussed above ask something like:

Here is the existing worker, the existing job and the existing workflow. What happens if we add AI?

That is a reasonable first experiment. It may not be the mature system. The larger change may come when the work itself is redesigned.

A future workflow may distribute skills differently, turn full-time roles into occasional oversight, or separate search, verification, judgement and accountability in ways that today's jobs do not.

We cannot measure that future directly, but we should at least avoid treating today's job architecture as a law of nature.

**These experiments test AI inside inherited work.**

That is another reason the workflow — not "human" or "AI" in the abstract — should remain the unit of analysis.

## **Implementation status at Hedegreen Research**

---

This article proposes a method for evaluating human–AI newsroom workflows. It is not a benchmark of one specific model and should not be read as a permanent claim about "AI performance" in general.

The cited studies report results for the models, tasks, interfaces and users they tested. Their publication dates do not make those results a benchmark of current models. Older studies can still help explain mechanisms.

Human review alone is not treated as sufficient; verification must still return to preserved source evidence.

## Four principles

---

A minimal standard for journalism could be:

1. **Compare workflows on the actual task.** Do not compare “humans” and “AI” in the abstract.
2. **Measure the errors that matter.** Include severity, detectability and propagation, not only average accuracy.
3. **Test independence and complementarity.** Agreement is not verification. Ask whether one system can actually catch the other's mistakes.
4. **Verify the combined workflow.** Do not infer human–AI performance from human-only and AI-only scores.

These rules do not tell a newsroom to use AI. They tell it how to think before deciding.

## What remains unresolved

---

There are still difficult questions. How should a newsroom weight catastrophic but rare failures? How independent does a human reviewer need to be? When should the AI speak first, second or not at all? How much does expertise change the result? How quickly should old human–AI studies be considered stale? And how much of today's evidence will survive once the jobs themselves change around the technology?

Those are not reasons to stop testing. They are reasons to test the right object.

The interesting question is not:

Does AI make mistakes?

Of course it does. The interesting question is:

**Which workflow leaves fewer important mistakes uncaught?**

— Dennis Hedegreen, trying to see the structure

## Sources

---

1. Ngo (2025), Human–AI collaboration in high-stakes decisions: a meta-analysis of healthcare and public sectors (<https://www.tandfonline.com/doi/full/10.1080/13504851.2025.2586160>).
2. Dell’Acqua et al. (2026), The Cybernetic Teammate: A Field Experiment on Generative AI and Teamwork (<https://pubsonline.informs.org/doi/10.1287/orsc.2025.20702>).
3. Cui et al. (2026), The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers (<https://pubsonline.informs.org/doi/10.1287/mnsc.2025.00535>).
4. Rosenbacke et al. (2024), False conflict and false confirmation errors are crucial components of AI accuracy in medical decision making (<https://www.nature.com/articles/s41467-024-50952-3>).

5. Wies et al. (2024), Reply to: False conflict and false confirmation errors are crucial components of AI accuracy in medical decision making (<https://www.nature.com/articles/s41467-024-50954-1>).
  6. Gonzalez et al. (2026), Toward a science of human–AI teaming for decision making: A complementarity framework (<https://academic.oup.com/pnasnexus/article/5/3/pgag030/8490283>).
- 

#### RELATION MEMORY

NEARBY [Refusing AI Is Not a Journalistic Principle](#) · Previous instalment in Journalism, Reassembled. Article 008 separates principles from controls; Article 009 asks how a controlled human-AI workflow should be evaluated.

NEARBY [Sources Should Travel With Claims](#) · Article 005 establishes the preserved evidence relationship required when a human or AI surfaces, transforms or checks a claim.

NEARBY [Claims Need States](#) · Article 004 provides the claim-state discipline needed to distinguish surfaced material from verified evidence.

---

CANONICAL URL: [HTTPS://HEDEGREENRESEARCH.COM/ARTICLES/AI-MAKES-MISTAKES-IS-NOT-AN-ARGUMENT/](https://hedegreenresearch.com/articles/ai-makes-mistakes-is-not-an-argument/)