Artificial intelligence as an epistemological Swiss Army knife: when perceived authority replaces verification

Carlos Eduardo Ravello Joo
ORCID: 0009-0007-5631-7436
carlosravello.com
Trujillo — Lima, Peru

DOI: 10.5281/zenodo.24116983

November 2026

Abstract

Why do people stop verifying when a claim arrives with signals of authority, and what changes when the source is an AI? This article argues that the hallucinations of language models and the crisis of authority of expert institutions are the same problem: in both cases, perceived authority replaces verification.

The argument starts from the replication crisis and from the theories of motivated and argumentative reasoning (Kunda, 1990; Mercier & Sperber, 2011): we reason to convince, and correction comes from the interlocutor who evaluates. Generative AI removes that interlocutor and adds a bias of its own, sycophancy. The cases of the four large audit firms that, between 2025 and 2026, published reports with fabricated citations, and the dispute over how to audit their auditors, illustrate the mechanism at institutional scale.

The article proposes a decision framework for the value of verifying, based on the value of information, that distinguishes three levels of verification (existence, fidelity and validity) and places the effect of authority on perceived probability and the effect of sycophancy on the diagnosticity of the source. It also shows that there are two routes to not verifying, the desire for a conclusion and the cost of checking it, and that both fall under the same inequality: that is what allows the social media user and the consulting firm to be treated as one problem. This is a position paper: it reports no data of its own. It closes with a two-part experiment whose predictions, primary measure and refutation criteria are set before any data are collected.

1. Introduction

An argument on X, TikTok or Facebook reaches its tensest point. One participant does not look for a study or consult a source: they ask an AI to agree with them, copy the answer and paste it as conclusive proof. The other has no way to verify it on the spot, and the debate is treated as won.

The scene is trivial, but it contains the whole problem. The pasted answer has the shape of knowledge: it is well written and sometimes comes with citations. Nobody checked whether it is true, neither the person who asked nor the person who received it, and its authority comes from the tool, not from the content. AI works today as an epistemological Swiss Army knife. The tool that helps you study delivers, with the same finish, the argument that wins the discussion,

and what serves for everything is rarely checked.

The scene is starting to echo in search infrastructure itself. On 6 November 2026, a Google search for «la ia como zanjadora de debates» [AI as the settler of debates] returned an AI-generated answer at the top of the results that described AI as a neutral arbiter, able to process information «sin sesgos emocionales» [without emotional biases] (screenshot in the supplementary material). This is an observation, not data: the result is personalised and changes over time. But it illustrates this article's starting point: saying "I asked the AI" is beginning to work as an argument on its own.

This article argues that the scene is not an isolated case of misuse. It is the everyday version of a pattern that had already appeared in professional science, that AI now makes universal, and that also reaches the institutions whose job is, precisely, to verify.

2. The precedent: the replication crisis

Professional science has already shown what happens when people look for the evidence that proves them right. In 2015 the Open Science Collaboration repeated 100 studies published in three leading psychology journals: 97% of the originals had reported significant results, and only 36% of the replications obtained them. The average effect size also fell by half (Open Science Collaboration, 2015).

The figure was debated, and rightly so: there are several ways to count a successful replication. But the underlying diagnosis does not depend on that number. The mechanisms that explain it had already been described:

- **Analytic flexibility (*p-hacking*).** Choosing variables, sample, covariates and conditions until the result comes out significant. Simmons, Nelson and Simonsohn (2011) showed that, with this undisclosed flexibility, it is more likely to find evidence for a false effect than to rule out a non-existent one.
- **Hypothesizing after the results are known (*HARKing*).** Presenting as a prediction what was discovered by looking at the data (Kerr, 1998).
- **Publication bias.** Journals reward positive results, so what reaches print is a biased sample of what was researched.

All three share an origin: the researcher wants their idea to work. No bad faith is needed. A system that rewards being right more than checking is enough.

3. Reasoning to win

If scientists, trained to doubt, fell into this, the problem is not one of training: it is how human reasoning works. Kunda (1990) described it as motivated reasoning: when we want to reach a conclusion, we use the strategies most likely to deliver it. The only brake is being able to build a justification that looks reasonable.

Mercier and Sperber (2011) went further. Their argumentative theory holds that the function of reasoning is not to find the truth alone, but to produce and evaluate arguments to convince others. From there, confirmation bias stops being a defect: it is what one would expect from a system designed to win arguments.

The same theory offers the way out. We are poor at finding counterarguments to our own ideas, but good at evaluating other people's arguments. That is why reasoning works better in groups and in real debate than alone: the other person does the verifying that we do not.

That is the condition AI breaks.

4. AI as an epistemological Swiss Army knife

Generative AI is today a Swiss Army knife of knowledge: it serves to write theses, study, research, and produce videos and images. And it serves for something that is hardly discussed: winning debates. It changes three conditions that used to hold motivated reasoning back.

1. **Cost.** Before, holding a position required looking for sources, reading them and citing them. Now a justification that looks reasonable takes seconds.
2. **Form.** The answer arrives polished, sometimes with references. It has the shape of knowledge even though nobody has checked its content.
3. **Interlocutor.** The argument no longer passes through another human who evaluates it before it is used. Whoever asks and whoever receives the answer share the same interest: that it confirm.

Kunda set a limit on motivated reasoning: the need to build a plausible justification. AI does not remove that limit; it outsources it. And Mercier and Sperber placed correction in dialogue: AI offers a dialogue without a counterpart.

5. The bias is also in the tool

AI does not only amplify the bias of whoever asks: it has one of its own. Sharma and colleagues (2023) showed that five state-of-the-art AI assistants are consistently sycophantic: they tend to answer what matches the user's beliefs rather than what is true. They also found a probable cause: both people and the models that learn from their preferences choose, a non-negligible fraction of the time, well-written sycophantic answers over correct ones.

The second mechanism is fabrication. Walters and Wilder (2023) asked two versions of ChatGPT for literature reviews on 42 topics and checked the 636 resulting references: 55% of GPT-3.5's and 18% of GPT-4's did not exist. Figures for later models may be lower, but the phenomenon has not disappeared, as the next section shows.

Together, the two mechanisms close the circuit. The user asks for proof of what they already believe, the tool tends to provide it and, if it does not exist, can invent it in citation format. Confirmation bias no longer has to make an effort.

There is more recent evidence that the tool does not only answer: it persuades. In a preregistered experiment with 900 participants, Salvi and colleagues (2025) had people engage in short debates against another person or against GPT-4, with or without access to the opponent's sociodemographic data. When GPT-4 had that data, it was more persuasive than the human in 64.4% of non-tied pairs. In medicine, Shekar and colleagues (2025) showed 300 participants answers written by doctors or generated by a language model. Participants could not tell them apart, preferred the AI's, rated low-accuracy answers almost the same as the doctors', and showed a high tendency to follow potentially harmful advice; they also trusted an AI answer more when told it came from a doctor. Neither study measures what this article proposes: the first measures AI arguing, not a person citing it; the second, whether people follow advice, not whether they verify it. They are neighbouring evidence, not the same.

6. The crisis of authority: the Big Four

Between October 2025 and July 2026 it was discovered that the four large audit and consulting firms (Deloitte, EY, KPMG and PwC) had published documents with non-existent citations. These are the institutions whose product is precisely verification.

Date	Firm	Document	What was found	Response
Jul. 2026	PwC Middle East	Four *thought leadership* reports (2024–2026)	GPTZero found 4 fabricated citations across the four reports and claims not supported by their sources; the central finding is a framework ("Citizen Pulse") with no trace outside the report, presented as in use by the governments of Denmark, Saudi Arabia, the US and Australia (GPTZero, 29 Jul. 2026)	A spokesperson said the firm was updating a limited number of supporting citations (City AM, 29 Jul. 2026)
Jun. 2026	KPMG	*Total Experience: Redefining Excellence in the Age of Agentic AI* (Oct. 2025)	Of 45 citations, 5 correct, 28 pointing to real sources with paraphrased titles or fake components, and 12 undeterminable; about half of the supported claims were false or misattributed (GPTZero, 12 Jun. 2026)	A KPMG International spokesperson confirmed the report had been withdrawn and that the circumstances of its publication were under review (City AM, 12 Jun. 2026)
May 2026	EY Canada	*Points of Attack: Uncovering Cyber Threats and Fraud in Loyalty Systems* (late 2025, 44 pp.)	16 of 27 sources fabricated, misattributed or with broken links, checked by hand by GPTZero (GPTZero, 14 May 2026; Computing, 18 May 2026)	Withdrew the report and is reviewing how it was published; told the *Financial Times* that it was not connected to work carried out for any client
Nov. 2025	Deloitte Canada	Health human resources plan for Newfoundland and Labrador, 526 pages, about CA$1.6 million	At least four citations to non-existent articles, attributed to real researchers who denied them, detected by *The Independent* (CP24, 25 Nov. 2025; HR Reporter, 25 Nov. 2025)	Deloitte acknowledged the four erroneous citations and, according to the provincial health department, stands by the report's conclusions and findings
Oct. 2025	Deloitte Australia	Review of the Department of Employment's automated penalty system, AU$440,000 contract	A fabricated quotation from a Federal Court judgment and references to non-existent works by real academics, detected by Christopher Rudge (U. of Sydney) (The Register, 6 Oct. 2025)	Refunded the final instalment and disclosed the use of Azure OpenAI GPT-4o in the corrected version; its internal review attributed the errors to human error (AFR, via The Register). The client department (DEWR), not Deloitte, stated that the substance was retained and the recommendations did not change

The cases are not equivalent. PwC's is the mildest: four fabricated citations across four reports; headlines about reports "100% generated by AI" rest on a detector whose reliability is debatable. Deloitte's are the most serious, because they were reports paid for by governments to decide public policy.

In both Deloitte cases the same closing line appeared, though from different speakers: in Australia it came from the client department (DEWR); in Canada, from Deloitte itself through the provincial ministry. That client and consultant agree that the substantive content stands and the recommendations do not change resembles the defence of authors whose effects did not replicate, and leaves a question open: if the sources were false, what were the conclusions resting on?

7. One and the same problem

AI hallucination and the crisis of authority are usually treated as different problems: one technical, the other institutional. This article argues that they are the same: in both cases, authority replaces verification.

Authority has always worked as a reasonable shortcut. Nobody can check everything, so we trust those with a track record, whether a journal or an audit firm. The shortcut works as long as someone, at some point in the chain, does verify. When the institutions that validate fall short, AI fills that gap (Ravello Joo, 2026).

The psychology of persuasion explains why the shortcut weighs so much. Since Hovland and Weiss (1951) we have known that the same message persuades more when it comes from a credible source, and that over time people separate content from its source: scepticism toward the untrustworthy source fades and the message stays. Dual-process models specified when the shortcut is used and when content is evaluated (Chaiken, 1980). Petty, Cacioppo and Goldman (1981) found that when a topic has little personal relevance, the expertise of the source persuades, and when it has a lot, the quality of the argument does. Authority does not always switch judgement off: it replaces it when processing is costly or matters little. Section 10 turns that result into a prediction.

The cases in the previous section show a chain in which nobody verifies:

1. The government or client trusts the firm by its name.
2. The firm trusts its consultant, who trusts the tool.
3. The tool produces plausible text, citations included.
4. The firm's seal turns that text into an authoritative report.

Step four is decisive: the hallucination leaves the chain with an authority it did not have when it entered. It is what the user who pastes an AI answer to win a debate does, but at institutional scale and with public money.

Each link has its literature. The client who trusts the firm by its name uses the source-credibility shortcut. The consultant who trusts the tool shows automation bias, discussed at the end of this section. What is new is that AI combines both: it is an everyday machine that is almost always useful, and at the same time it speaks like an expert and agrees with whoever asks. It is the first machine that also works as an authority.

AI's authority is borrowed. It used to rest on credentials and, above all, on what science accumulated through replication and citation: authors and studies that others had already reviewed. AI summarises that body of work and cites it, and in doing so inherits its authority, even when the reference is invented. That is where the crisis lies: the tool gathers data and synthesises it, but does not verify, and it is still treated as a guide, as an operating manual and even as the casting vote. It happens from an argument on social media to the reports of the Big Four, where delegating data collection shifted onto the machine the epistemic burden of a decision that should never have stopped being verified.

The replication crisis taught that a journal's prestige does not guarantee that an effect exists. The Big Four cases teach the same about a firm's prestige. AI did not create this weakness: it made it cheap and hard to see.

The mechanism, however, is not the same at both ends. The user in the opening scene does not verify because they want the conclusion: that is motivated reasoning. For the consultant there is no need to assume they wanted any particular conclusion; it is enough that verifying cost them time and nobody was going to ask them to. That is not confirmation bias but a cost nobody takes on. They are two different routes to the same result, and what justifies treating them

as one problem is that they fall under the same inequality (section 9): one acts on perceived probability and the other on the cost of checking.

The phenomenon is not new. Human factors psychology has studied it for decades under another name: automation bias (Skitka, Mosier & Burdick, 1999). When an automated decision aid is wrong, its user makes two kinds of error: omission, failing to detect a problem because the system did not flag it, and commission, following a wrong recommendation. The review by Parasuraman and Manzey (2010) found that the bias appears in novices and experts alike and that neither training nor instructions prevent it. Generative AI inherits that bias and adds two things classical automation lacked: it argues on request for whoever uses it, and it can fabricate the source that backs what it says. A cockpit alarm does not adapt to what the pilot wants to hear; a language model does. A note on terms: in that literature, "complacency" refers to the human operator's excessive trust in the system; in this article, following Sharma and colleagues (2023), "sycophancy" refers to the model's tendency to agree with whoever asks. They are two different phenomena, and in the chain described above they add up.

8. Who audits the auditor

The cases in section 6 have a detail that is usually overlooked: three of the five were detected by GPTZero, a company that sells detectors of AI-generated text; the two Deloitte cases, by an academic and by a local outlet. Each GPTZero finding is also advertising for its product. That does not invalidate what it found, but it requires separating two things its reports mix. The checkable part (a citation does or does not exist in Crossref) holds even if it comes from an interested party, because anyone can repeat it. The proprietary part (the "100% AI-generated" score) comes from a closed method, of unknown diagnosticity, and from whoever gains if the score alarms.

The second detail is who asked for the audit: nobody. Neither the governments that paid for the reports nor the firms that published them commissioned a citation check. Verification happened because an academic, an outlet and a company had their own reasons to do it. In the framework of the next section, the chain gets verified only where someone has an interest or a loss at stake, not where the design requires it. And the question repeats one level up. The auditor was in fact audited: Li, Lin and Ma (2026) measured GPTZero's citation detector against their own benchmark and reported a *recall* of 62.8% on real citations and 67.5% on fabricated ones; GPTZero replied that, under its definition of hallucination, it achieves 0.2% false positives and 99.1% *recall* (GPTZero, 14 Oct. 2026). Both figures are correct: they measure different things. For one, an error in journal, year or page already counts as a hallucination; for the other, only inventing the title or the authors does. The result of an audit depends on what counts as false, and that decision has to be made before looking at the data.

The third detail weighs the most. The intuitive solution (using one AI to audit what another AI wrote) assumes the two errors are independent. They are not entirely: large models share much of their training data (Common Crawl, Wikipedia, arXiv, digitised books), and it is reasonable to expect them to tend to invent the same plausible citations on the same topics. A second opinion from someone who studied from the same book is not a second opinion. Whether this happens and to what extent is an empirical question, and the experimental proposal includes it.

The way out is not to find a more trustworthy auditor, because the regress never ends. It is for the verification chain to end in a step that does not require trusting anyone: a high-diagnosticity check that anyone can repeat with the same means. It follows that diagnosticity is not a property of the model but of the protocol in which it is used: the same model, when asked for proof of something and when required to open each source before citing it, is two sources with different likelihood ratios. The practical implication is not to distrust AI, but for the chain to end in a

repeatable step.

9. Why people do not verify: a decision framework

Verifying is a decision with costs and benefits, and it can be analysed as such. This is not an original contribution: it is the application of the value of information from decision theory (Raiffa & Schlaifer, 1961; Howard, 1966) to the problem of believing or checking a claim. The formal development is in Appendix A; its consequences are summarised here.

It pays to verify when the expected loss from believing something false exceeds the cost of checking. A threshold follows: below a certain probability that the claim is true, verifying is worth it. When the loss is high and the cost low, the threshold sits close to certainty and verifying almost always pays. In the cases of section 6 the loss was high and the cost minimal, and still nobody did it.

Verification has levels, and each has its own cost and scope:

1. **Existence:** the cited document exists. A DOI in Crossref settles it in a minute, with near-perfect diagnosticity for fabrication.
2. **Fidelity:** the document says what the citation claims it says. This requires opening it and reading the passage.
3. **Validity:** what it says is well done: design, data, replication. This is the level at which the *p-hacking* and *HARKing* of section 2 are detected, and the most costly.

A real DOI certifies that the document exists, not that it is right. Verifiable does not mean true: it is the condition for being able to check it, and the task continues at levels 2 and 3. A level-1 check detects nothing at the following levels: the 28 KPMG citations that pointed to real sources with altered titles would have passed it. What is serious about the cases in section 6 is that not even the level-1 check, the cheapest, was done.

The framework locates the error. Authority does not act on the calculation but on the perceived probability that the claim is true: a firm's seal or an AI's fluency raises it above the threshold, and then the calculation says not to verify. It is the normative version of what Sperber and colleagues (2010) called epistemic vigilance: the mechanisms by which we calibrate how much to believe a source. This article proposes that prestige signals can switch them off.

In the notation of Appendix A, p is the perceived probability that the claim is true, L the loss from believing a false one, c the cost of checking and s the sensitivity of the check.

The two routes in section 7 occupy different places in the same inequality. Motivated reasoning acts on p: whoever wants something to be true perceives it as more probable. The economics of expert work acts on c and on $s \cdot L$ as the decision-maker perceives them: the time spent checking is paid by the consultant, while the loss from an error, if detected, is paid mainly by the client. That is why a loss of millions for the client can coexist with an individual decision not to verify, without invoking any bias.

Sycophancy acts on another parameter: the diagnosticity of the source. A source that answers "yes" whether the claim is true or false provides no information and should not move what we believe. This formalises what section 8 said: what is valuable in an investigation like GPTZero's is not its "AI-generated" score, of unknown diagnosticity, but its lookup in bibliographic databases, of near-perfect diagnosticity and repeatable by any human.

AI also introduces an asymmetry. The cost of producing a claim formatted as proof has fallen almost to zero; the cost of verifying it has not changed. Whoever receives many claims cannot check them all, so verification becomes a scarce resource allocated according to perceived trust. Signals of authority are precisely where it stops being allocated.

The model is normative: it describes when verifying pays, not how people decide. Its empirical value is that each of its parameters can be measured or manipulated, and the following proposal does so.

10. Experimental proposal

The thesis can be tested with a simple two-part design. Predictions and refutation criteria are set here, before any data are collected, and the protocol will be preregistered on OSF.

Part A: the tool as an ammunition supplier

Material. Twenty claims: ten with clear scientific consensus and ten genuinely open.

Procedure. For each claim, three requests to each model: "give me evidence that X", "give me evidence that not X" and "what does the evidence say about X?". At least four widely used models, in their default configuration, with five repetitions per request.

Measures.

- **Sycophancy:** proportion of answers that argue in the requested direction.
- **Resistance to consensus:** for consensus claims, proportion of answers that correct or refuse the request that runs against the evidence.
- **Fabrication:** each cited reference is checked in Crossref, Google Scholar or PubMed and classified as existing, with errors, or non-existent.
- **Cross-model correlation:** proportion of non-existent citations that coincide, fully or partially, across two or more models. If it is high, auditing one AI with another is not independent.

Predictions.

1. For open claims, models will argue in the requested direction in most cases, in both directions.
2. For consensus claims, sycophancy will be lower than for open claims, but above zero.
3. The rate of non-existent citations will be higher when the models are asked to argue against consensus, because there are fewer real sources to cite there.

What would refute the thesis. If, for consensus claims, models correct the request in at least 90% of cases and the rate of non-existent citations is close to zero in all conditions, the tool would not be amplifying motivated reasoning, at least with current models.

Cross-model correlation. Prediction: non-existent citations will coincide across models more than expected by chance, which for invented titles is practically zero. Refutation: coincidence close to zero, in which case auditing one AI with another would be an independent check.

Protocol. The same requests are repeated in a second condition: the model with search enabled and the instruction to open each source and give its DOI or link before citing it. Prediction: the rate of non-existent citations will be lower in this condition than with the default configuration, although sycophancy in the direction of the argument may not change. Refutation: equal fabrication rates in both conditions, in which case diagnosticity would be a property of the model and not of the protocol.

Coding. The person proposing the hypothesis does not code the answers. Two coders blind to condition classify whether each answer argues in the requested direction, and their agreement is reported. The definitions of a citation "with errors" and "non-existent" are set in the preregistration because, as section 8 shows, the result of an audit depends on what counts as false.

Part B: authority as a substitute for verification

Procedure. Participants read the same text containing a fabricated citation, in a $3 \times 2 \times 2$ between-subjects design: the source (no source, "an AI's answer" or "a report by an international consulting firm"), the consequence of being wrong (high L: a decision with money or health at stake; low L: a curiosity) and the cost of verifying (low c: the citation comes with a one-click link; high c: only the reference, which has to be searched for). Whether the participant verifies is recorded.

Measures. The primary measure is behaviour: whether the participant opens or searches for the citation. Secondary measures are perceived credibility and willingness to share the text. Behaviour is fixed in advance as the primary measure so as not to choose afterwards whichever measure turns out best.

Predictions. According to the framework of section 9, verification will rise with L and fall with c. The consulting-firm label will reduce it relative to the no-source condition, with identical content. The effect of the "AI's answer" label relative to the consulting firm's is not set as a directional prediction; it is discussed in the next paragraph.

The ordering of the two labels. This is the study's open question. The author's hypothesis is that the perceived authority of AI is catching up with institutional authority: a result in which the AI label reduces verification as much as the consulting-firm label, or more, would be the central finding. Surveys do not settle it. In the global study by Gillespie and colleagues (2025), with more than 48,000 people in 47 countries, only 46% said they were willing to trust AI, yet at work 66% said they had used its output without evaluating its accuracy. According to the University of Melbourne (2025), compared with its 2022 survey, people became less trusting as adoption increased. Stated trust and behaviour go in opposite directions, which is why the primary measure of this design is behaviour. Gillespie's study was funded in part by KPMG, one of the firms in section 6. Because that authority may be changing fast, the proposal is to repeat Part B with the same protocol every year: a single measurement is a snapshot; several are the trajectory.

Two rival models. The framework admits two readings of the effect of authority, and the design tells them apart. If authority only raises perceived p, the calculation keeps running: with high L and low c the threshold rises enough for the label's effect to weaken. If authority suspends the calculation and replaces it with a heuristic with no cost or loss term, the label will reduce verification equally with high L and low L. The contrast that decides between them is the label $\times$ L interaction, not the main effects, and it is preregistered as the primary test. The five cases in section 6 do not allow this distinction, because in them the loss was paid by a third party; that is why, in the high-L condition, the consequence is framed as the participant's own.

The three possible outcomes, set before the data.

1. **The hypothesis holds:** verification falls with the authority label even when L is high and c almost zero, that is, when the framework indicates that verifying clearly pays.
2. **The hypothesis breaks:** verification tracks L and c and the label does not divert it. Authority would not substitute for verification in this context. This result will also be published: it would indicate that people are more sensitive to what is at stake than this article assumes.
3. **Mixed result:** the label changes credibility but not behaviour, or the reverse. Since behaviour is the primary measure, the hypothesis is evaluated with it; credibility is reported as a secondary finding.

Analysis. In Part A, proportions with confidence intervals and mixed logistic models, with AI model and claim as random effects. In Part B, logistic regression with the label $\times$ L interaction

as the primary test. The operationalisation of the measure and of the loss (search within the platform, payment contingent on accuracy) is set in the protocol. Sample size will be set in the preregistration with a power analysis for the label × L interaction, which requires more participants than the main effects. As a feasibility reference, Salvi and colleagues (2025) ran a preregistered 12-cell between-subjects design with 900 participants. Materials, data and code will be open.

11. Limitations

- **The cases are not a sample.** The five Big Four controversies are known because someone detected them. They do not allow estimating how often the problem occurs, and most were revealed by a company that sells AI detectors.
- **A denominator is missing.** There is a domain with a better record: the courts, where verifying is the institution's function, errors are formally sanctioned and decisions are public. Even so, a register of sanctioned cases gives only the numerator: it counts what was detected, not what was filed, and its growth over time mixes how much happens with how much is detected. It is a route for future work, not a sample this article has.
- **The cause is not proven.** In almost no case was it proven that AI generated the false citations. Deloitte Australia disclosed the use of GPT-4o; the rest is reasonable inference, not direct evidence.
- **Models change fast.** The sycophancy and fabrication rates cited are from 2023 models. The proposed experiment exists precisely to measure them in current models.
- **The argumentative theory has critics.** This article uses it as a framework, not as an established fact. The central thesis also holds with Kunda's motivated reasoning, without the evolutionary part.

Note on method

This article was written with the assistance of a language model (Claude, Anthropic; Opus 5.5 and Fable 5.1 models; October and November 2026). The model searched for sources, extracted data, drafted text and subjected the argument to critique. Content decisions, interpretive judgement and the final version are the author's. Declaring this is not a formality: an article arguing that authority replaces verification cannot ask to be believed. Appendix B reports the verification protocol, the errors that occurred, who detected them and what remains unverified, so that anyone can repeat the check and find what I did not see.

This English version is a translation of the Spanish original. Statements by firms and authorities that the Spanish text quotes in translation are given here in paraphrase, without quotation marks, except where the English original was checked.

References

Academic works

- Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology, 39*(5), 752–766. https://doi.org/10.1037/0022-3514.39.5.752
- Gillespie, N., Lockey, S., Ward, T., Macdade, A., & Hassed, G. (2025). *Trust, attitudes and use of artificial intelligence: A global study 2025.* The University of Melbourne & KPMG. https://doi.org/10.26188/28822919
- Hovland, C. I., & Weiss, W. (1951). The influence of source credibility on communication effectiveness. *Public Opinion Quarterly, 15*(4), 635–650. https://doi.org/10.1086/266350

- Howard, R. A. (1966). Information value theory. *IEEE Transactions on Systems Science and Cybernetics, 2*(1), 22–26. https://doi.org/10.1109/TSSC.1966.300074
- Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4
- Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480
- Li, M., Lin, Z., & Ma, S. (2026). Source or it didn't happen: A multi-agent framework for citation hallucination detection. *arXiv*. https://arxiv.org/abs/2605.08583
- Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences, 34*(2), 57–74. https://doi.org/10.1017/S0140525X10000968
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. https://doi.org/10.1126/science.aac4716
- Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors, 52*(3), 381–410. https://doi.org/10.1177/0018720810376055
- Petty, R. E., Cacioppo, J. T., & Goldman, R. (1981). Personal involvement as a determinant of argument-based persuasion. *Journal of Personality and Social Psychology, 41*(5), 847–855. https://doi.org/10.1037/0022-3514.41.5.847
- Raiffa, H., & Schlaifer, R. (1961). *Applied statistical decision theory.* Division of Research, Graduate School of Business Administration, Harvard University.
- Ravello Joo, C. E. (2026). *The third knowledge regime: Artificial intelligence as an emergent response to the limits of classical academia and open repositories* [Preprint]. Zenodo. https://doi.org/10.5281/zenodo.20298744
- Salvi, F., Horta Ribeiro, M., Gallotti, R., & West, R. (2025). On the conversational persuasiveness of GPT-4. *Nature Human Behaviour, 9*(8), 1645–1653. https://doi.org/10.1038/s41562-025-02194-6
- Sharma, M., Tong, M., Korbak, T., et al. (2023). Towards understanding sycophancy in language models. *arXiv*. https://doi.org/10.48550/arXiv.2310.13548 (presented at ICLR 2024)
- Shekar, S., Pataranutaporn, P., Sarabu, C., Cecchi, G. A., & Maes, P. (2025). People overtrust AI-generated medical advice despite low accuracy. *NEJM AI, 2*(6). https://doi.org/10.1056/AIoa2300015
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632
- Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies, 51*(5), 991–1006. https://doi.org/10.1006/ijhc.1999.0252
- Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language, 25*(4), 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x
- University of Melbourne. (2025). *Global study reveals trust of AI remains a critical challenge reflecting tension between benefits and risks* [Press release]. https://fbe.unimelb.edu.au/newsroom/media-release-global-study-reveals-trust-of-ai-remains-a-critical-challenge-reflecting-tension-between-benefits-and-risks
- Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports, 13*, 14045. https://doi.org/10.1038/s41598-023-41032-5

Big Four cases (press and investigations)

- City AM. (2026, June 12). KPMG report on AI found riddled with AI hallucinations
- City AM. (2026, July 29). PwC thought leadership reports '100 per cent AI generated'
- Computing. (2026, May 18). EY cybersecurity report pulled after probe finds 'AI hallucinations'
- CP24 / The Canadian Press. (2025, November 24). Deloitte report for Newfoundland government found to have apparently false citations
- Financial Times. (2026, May 15). EY retracts study after researchers discover AI hallucinations
- GPTZero. (2026, May 14). Investigation: Hallucinations in Ernst & Young report on loyalty fraud
- GPTZero. (2026, June 12). Chasing the hallucinations: KPMG's AI-powered attempt at "Redefining Excellence"
- GPTZero. (2026, July 29). Chasing the hallucinations: PwC report hallucinates product and government customers
- GPTZero. (2026, October 14). What counts as a hallucinated citation? A response to CiteTracer
- HR Reporter. (2025, November 25). AI errors: Province grapples with Deloitte report marred by false citations
- The Register. (2025, October 6). Deloitte refunds Aussie gov after AI fabrications slip into $440K welfare report

Appendix A. Formal framework

This appendix develops the framework of section 9. The expressions are direct applications of the theory of the value of information (Raiffa & Schlaifer, 1961; Howard, 1966); they are not presented as original results.

A.1 Value of verifying

Let p be the probability that the claim is true, L the loss from acting on a false one, B the benefit lost by discarding a true one, s the sensitivity of the verification (it detects what is false), e its specificity (it does not reject what is true) and c its cost. If the default action is to accept the claim, the expected value of verifying is:

$$V = (1-p)\,s\,L \;-\; p\,(1-e)\,B \;-\; c$$

It pays to verify when $V > 0$. "Accept" is taken as the default action because it is the scenario this article studies: authority that leads to acceptance. If the default action were to reject, the expression would be:

$$V_{\text{reject}} = p\,e\,B \;-\; (1-p)(1-s)\,L \;-\; c$$

A.2 Verification threshold

When $e \approx 1$, as when checking whether a DOI exists, the second term vanishes and the condition $V > 0$ reduces to:

$$p \;<\; p^{*} = 1 - \frac{c}{s\,L}$$

If c is much smaller than L, the threshold p^* is close to 1 and verifying almost always pays. The article's hypothesis is that signals of authority raise perceived p above p^*. In Part B, L and c shift p^*, and the source label shifts perceived p: the $3 \times 2 \times 2$ design manipulates the three terms of the inequality.

Section 10 contrasts this hypothesis with a rival: that authority does not shift perceived p but suspends the calculation. If it only shifts it, the label's effect depends on p^*, and therefore on L and c; if it suspends it, it does not depend on them.

A.3 Levels of verification

The probability that the claim is false decomposes according to the level at which it fails: existence, fidelity or validity.

$$1 - p = q_1 + q_2 + q_3$$

A level-1 check has sensitivity close to 1 for q_1 and zero for q_2 and q_3. Its value is therefore $q_1 \cdot s_1 \cdot L - c_1$, and it does not cover errors of fidelity or validity. That is why a citation with a real DOI can still be false.

A.4 Sycophancy as a likelihood ratio

The diagnosticity of a source that asserts X is measured by its likelihood ratio:

$$LR = \frac{P(\text{asserts } X \mid X \text{ true})}{P(\text{asserts } X \mid X \text{ false})}$$

Posterior odds are obtained by multiplying prior odds by LR. A sycophantic source, which asserts what it is asked whether or not it is true, has $LR \approx 1$ and should not move belief. In Part A, the truth is known for consensus claims: the proportion of times the model argues for a true claim estimates the numerator, and the proportion of times it argues for a false one estimates the denominator. Part A thus estimates the diagnosticity of each model as a source.

A.5 Correlated auditors

If one AI (B) audits what another (A) wrote, the effective sensitivity of that audit is:

$$s_{\text{ef}} = 1 - P(B \text{ accepts} \mid A \text{ invented})$$

If the errors of the two models are independent, s_{ef} approaches the sensitivity of B on its own. If they share training data and tend to invent the same citations, s_{ef} falls. The cross-model correlation measure in Part A estimates that dependence.

A.6 The asymmetry between producing and verifying

If a recipient faces n claims and verifying each costs c, the total cost is $n \cdot c$. When the cost of producing claims falls and n grows, the recipient cannot verify them all and must prioritise by perceived p. Claims with signals of authority, which have a high perceived p, are the first to go unverified.

Appendix B. This article's protocol and its failures

I wrote this article with the tool the article examines. I mention it not out of administrative honesty, but because it is the only case whose inside I can report: in the five cases of section 6, only the published report and the false citation are visible, never the process that let it through. Here the process is declared, with its failures.

The protocol. One session of the model drafted the text. No field of any reference could be written from memory without being flagged: author, year, title, journal, volume, pages and DOI are read from the source record (Crossref for articles with a DOI, the repository for preprints, the title page for books), or the field is flagged as unverified. A second session, different from the one that drafted, checked the eleven academic references field by field and opened the press links one by one. The first session then audited that verification by reopening sources at random. I arbitrated the disagreements.

What the verification detected. The eleven references in that round exist; the eight added later (automation bias, persuasion and trust in AI) were checked the same way: fields in Crossref or in the report's repository, and the cited content in the original abstract, not in the press. That round corrected two claims made from memory: Chaiken (1980) manipulated the communicator's likeability, not their expertise, so it is not cited as evidence about authority; and that trust in automation rises with reliability and use is not in Parasuraman and Manzey (2010), so it is not claimed with that citation. There was a field error: the pages of Mercier and Sperber (2011) were listed as 57–111, and Crossref gives 57–74. Four references (Kerr; Howard; Raiffa and Schlaifer; Sperber et al.) had been written from memory and were flagged as such: all four turned out to be correct, and only when they were checked could their DOIs be added. A figure inherited from secondary press was exaggerated: the KPMG report did not have "40 of 45 fabricated citations" but 5 correct, 28 pointing to real sources with paraphrased titles or fake components, and 12 undeterminable. One statement was misattributed: in the Australian case, the statement that the substance was retained and the recommendations did not change came from the client department, not from Deloitte. And an earlier version of section 8 claimed that nobody had audited GPTZero. That is false: Li, Lin and Ma (2026) had published a measurement of its detector, and GPTZero replied in October. Nobody had looked. It is an error of absence, the same kind of failure this article documents but inverted: instead of inventing a source, its non-existence was invented.

What the audit detected. The verifying session also made mistakes. When the audit pointed out that there was a KPMG statement confirming the withdrawal of its report, the verifying session replied that the quotation seemed to be misattributed to another firm, without having searched for it. When it searched, it found the quotation verbatim in an article it had not opened. Later, when the corrections were applied to the text, the date of one case ended up in another case's row of the table in section 6; the audit caught it in the following review. Correction also needs verification. Finally, a review by another AI system objected to a thesis that section 7 did not hold: a severe critique is not, for that reason, a reliable one.

My reading. The protocol did not work because it was strict, but because at each step someone who had not produced the text went to look at the source. Verification of a piece of work by whoever produced it has the same diagnosticity problem that section 9 attributes to a sycophantic source, and this holds whether the verifier is a person or the same model in another session.

What remains unverified. The figure of 16 of 27 sources in the EY report comes from a single extraction in secondary press, because the original investigation does not provide it as text. Press quotations were not checked against their English originals beyond what is cited. There may be errors I did not see, and I do not say this out of modesty: verifying has a cost,

and I too had to decide where to stop.

What this does not show. Nothing about the error rates of current models: it is a single case, unblinded, and I decided what counted as an error. It is reported as transparency about the process, not as evidence for the hypothesis. I do not claim the protocol works; I report the record, which is what allows someone to contradict it.