When the Tool Agrees with You: Large Language Model Sycophancy, Examiner Impartiality, and the Admissibility of AI-Assisted Forensic Opinion

Devharsh Trivedi*

City University of New York, NY, USA

Abstract: Introduction. In *Matter of Weber* (2024) a surrogate's court put its own Microsoft Copilot query to three of its computers and got three different figures; the expert whose calculation prompted the exercise could not recall what prompt he had used. A model's output is a function of the prompt, which in forensic practice carries the examiner's hypothesis. Work on bias in digital forensics models the examiner deferring to the machine. Sycophancy inverts that: a model conforming to stated beliefs returns the examiner's hypothesis in the register of an independent instrument, manufacturing the appearance of corroboration.

Methods. We surveyed reliability screening for expert evidence in six regimes and tested the hypothesis. Six dated model snapshots answered five forensic interpretation items under three framings differing only in what the examiner asserts, sampled thirty times per cell for 2,700 trials. Replies were classified by a model judge validated against hand coding of 77 replies (Gwet's $\mathrm{AC}_1 = 0.855$).

Results. The hypothesis was not supported. Across 2,160 analysed trials, models asserted the incorrect proposition in 36.1% of trials when asked neutrally and 19.6% when an examiner asserted it; thirteen of twenty-four cells moved away from the examiner and ten did not move; the single cell that moved toward it had a baseline error of 93.3%, leaving 6.7 points of headroom. A second finding was sharper: on an item whose conclusion rested on an unsupported premise, one of six models contested it in 80% of trials and the other five in none of 150.

Discussion. Comparative screening shows an inversion: jurisdictions with the strongest reliability gates have no rule on expert use of generative AI, while Australia, which refuses a reliability gate, has the only binding instruments. Rule 702(d) is a sharper hook than *Daubert*'s error-rate factor, since a prompt-dependent error rate is not a property of a method. These models verify conclusions without auditing the reasons given, so disclosure built around the conclusion alone will not surface the failure.

Keywords: Digital forensics; expert evidence admissibility; large language models; sycophancy; cognitive bias.

1. INTRODUCTION

An expert retained in a New York estate matter used Microsoft Copilot to cross-check his damages calculation. Asked about it in court, he could not recall what prompt he had entered, could not say what sources the system had drawn on, and could not explain how it worked. The Surrogate's Court then did something that most courts confronted with an AI tool have not done: unable to reconstruct the expert's prompt, it composed a query of its own and put it to three of its computers. It received $949,070.97, $948,209.63, and, on the third machine, "a little more than $951,000.00" [1].

The court concluded that "prior to evidence being introduced which has been generated by an artificial intelligence product or system, counsel has an affirmative duty to disclose the use of artificial intelligence and the evidence sought to be admitted should properly be subject to a *Frye* hearing prior to its admission" [1].

Weber is commonly cited for the proposition that generative AI is unreliable and that its use should be disclosed. Read more closely it demonstrates something more specific. Three runs of one query produced three answers. The output was not a stable property of the tool. It was a function of the interaction, and the part of the interaction the expert could not reproduce was his own prompt.

This paper takes that observation seriously in the setting where it matters most. Digital forensic examiners are beginning to use large language models for triage, artifact interpretation, and report drafting. When they do, the prompt is not a neutral query. It carries the examiner's working hypothesis, because that is how practitioners actually ask questions: not "what does this artifact indicate" but "this shows the file was accessed on the fourteenth, correct?"

Large language models are known to be sycophantic: they conform their answers to beliefs the user has stated, agreeing with incorrect and subjective inputs in order to please [2,3]. In a forensic workflow this produces a failure mode that the existing literature on bias in digital forensics does not describe, and that no evidentiary rule anywhere currently addresses.

*Address correspondence to this author at the City University of New York, NY, USA; E-mail: DEVHARSH.TRIVEDI86@lagcc.cuny.edu

Contributions

1. We identify *hypothesis conformity* as a distinct threat to examiner impartiality, and distinguish it from automation bias, which runs in the opposite direction (Section 4).

2. We survey reliability screening for expert evidence across the United States, England and Wales, Australia, Canada, the European Union, and India, and report an inversion: strong reliability gates coexist with no AI-specific rule, and the one jurisdiction that refuses a reliability gate has the only binding rules on expert AI use (Section 5).

3. We argue that a sycophantic tool has a *prompt-dependent* error rate, which *Daubert*'s error-rate factor cannot accommodate, and that Rule 702(d) as amended in December 2023 is the better hook because it asks about *reliable application* rather than about the method in the abstract (Section 6).

4. We give a measurement protocol requiring no human participants, because the unit of analysis is the model rather than the examiner (Section 7), and run it over 2,700 trials on six deployed models (Section 8).

5. We report that the conformity hypothesis is not supported under stated-hypothesis framing, and that a different failure appears in its place: these models verify an asserted conclusion without auditing the reasoning offered for it. We propose disclosure requirements modelled on the Queensland practice direction, and withdraw a proposal our own data contradict (Section 9).

2. COGNITIVE AND CONTEXTUAL BIAS IN DIGITAL FORENSIC EXAMINATION

Forensic science has spent two decades confronting the finding that expert conclusions move with context. Dror, Charlton and Péron showed that fingerprint examiners presented with the same latent prints reached different conclusions when given extraneous contextual information [4], a result later extended to inter- and intra-expert consistency and to DNA mixture interpretation. Dror subsequently organised the phenomenon into a taxonomy of eight sources of bias [5] and a hierarchy of expert performance [6]. Kassin, Dror and Kukucka named the mechanism and traced its operation through the investigative process [7]. Both the National Research Council [8] and the President's Council of Advisors on Science and Technology [9] treated cognitive bias as a structural weakness of forensic practice rather than an occasional lapse.

Digital forensics received this literature later. Sunde and Dror set out the problem for the discipline [10] and then supplied the empirical anchor, applying the hierarchy of expert performance to fifty-three examiners and measuring both reliability and biasability [11]. Sunde subsequently proposed safeguards for examiner objectivity [12], and Renaud and colleagues offered a bias-neutralising framework for investigations [13]. Most recently Andersen, Sunde and Porter examined *tool-induced* bias, showing that how a tool presents data can itself bias analysis [14].

We are careful here about a claim that is easy to overstate. There is no established literature on *confirmation bias in digital forensics* as a named field. What exists is a literature on cognitive and contextual bias in digital forensics, within which confirmation bias figures as one component, generally traced back through Kassin *et al.* to the general psychological literature. The distinction matters because the broader claim is the accurate one.

The direction of deference

Every strand above, and the AI-specific work that follows it, models the human as the party at risk of deferring. Hefetz surveys bias in forensic evidence as it moves from expert analysis to AI-based decision tools, and organises the possibilities as offloading, collaborative partnership, and subservient use [15]. The European Union's AI Act encodes the same directionality: Article 14(4)(b) requires that human oversight enable the overseer "to remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system (automation bias)" [16].

Automation bias is real and well described. It is not the problem this paper is about.

3. SYCOPHANCY

Sycophancy is the tendency of a language model to favour beliefs the user has stated, agreeing with incorrect or subjective inputs rather than contradicting them. Perez *et al.* found the behaviour with model-written evaluations across model scales [3]; Sharma *et al.* characterised it across five production assistants and traced it to preference data in which human raters reward responses that match their own views [2].

Three features of the phenomenon matter for what follows.

It is prevalent rather than marginal. Benchmarks designed specifically to elicit it find sycophantic behaviour in a majority of trials.

It intensifies under authority. Where the user is positioned as an authority, follow rates rise sharply. In a forensic workflow the examiner is precisely such an authority: a domain professional, asserting a conclusion, within her own field.

It is sensitive to framing rather than to content. Presenting identical material as a question rather than as an assertion substantially reduces sycophantic agreement. This is what makes the phenomenon tractable as a protocol problem, and it is the basis of the mitigation we propose in Section 9.

Notably, a court has already described the behaviour without naming it. The Federal Court of Australia's practice note on generative AI lists among the technology's failure modes "confirmation that information is accurate if asked, even when it is not" [17]. That is sycophancy, in an instrument that governs expert evidence.

3.1. Sycophancy is not confirmation bias, and the difference is what makes it a forensic problem

The two are routinely conflated, and the conflation hides the risk. It is worth separating them on four axes.

Locus. Confirmation bias is a property of the examiner: a disposition to seek and weight evidence favouring a held hypothesis. Sycophancy is a property of an instrument the examiner consults. One is a state of mind, the other a transfer function from input to output.

Mechanism. Confirmation bias arises from motivated cognition and selective attention. Sycophancy arises from preference optimisation, in which raters reward agreeable responses, so the disposition is trained in and is a stable property of a weight snapshot rather than a lapse of discipline [2].

Observability. An examiner's confirmation bias leaves traces a peer reviewer can look for: which leads were pursued, which were dropped. Sycophancy leaves a reply that is fluent, hedged where a careful analyst would hedge, and formally indistinguishable from independent agreement. There is nothing in the output to inspect.

Correlation. This is the axis that matters. Two examiners with confirmation bias make errors that are at least partly independent, which is what makes peer review worth doing. A model snapshot conditioned on the examiner's stated hypothesis returns the same distribution of answers to every examiner who states that hypothesis. The error is common-mode across the profession and across cases, and it is invisible to precisely the mechanism, a second opinion, that forensic practice relies on to catch the first kind.

Why our result sharpens rather than dissolves this

The measured behaviour is not that models parrot whatever the examiner asserts. They do not: asserting an incorrect proposition moved every model away from it, not toward it (Section 8). What they do is endorse a *correct* assertion nearly without exception, with $\mathrm{agree}_T = 1.000$ for five of six snapshots, and they do so without auditing the reasoning that produced it.

That asymmetry is worse for forensic practice than uniform agreement would be, for a reason that is easy to miss. Uniform agreement is self-defeating: a tool that agrees with everything corroborates nothing, and practitioners would learn to discount it. Agreement that tracks correctness looks like a working instrument. An examiner who is right receives confirmation; an examiner who is wrong receives resistance. The examiner therefore has every reason to treat the model's assent as evidence, when in fact assent was never conditioned on the quality of their reasoning, only on the truth of their conclusion, which in a live case is exactly the thing not yet known. The instrument is reliable in the aggregate and uninformative in the individual case, and nothing on its face distinguishes the two.

4. THREAT MODEL: HYPOTHESIS CONFORMITY

Consider an examiner with a working hypothesis, which is the normal condition of forensic work rather than a defect in it. She consults a language model. The question she asks encodes what she already believes. The model agrees.

The output is not false in the way a hallucination is false. It may be entirely accurate as a restatement. What it is not is independent. The examiner has received her own hypothesis back, rendered in the voice of a separate instrument, and she has no way to distinguish that from corroboration.

We call this **hypothesis conformity**, and separate it from automation bias along three axes.

Table 1: Two failure modes of AI-assisted expert work, distinguished by the direction of deference.

	Automation bias	Hypothesis conformity
Direction of deference	Examiner defers to tool	Tool defers to examiner
Effect on the conclusion	Substitutes the tool's judgement	Amplifies the examiner's existing belief
Visible to the examiner?	Sometimes, on disagreement	No: agreement is the expected outcome
Detectable in the report?	Tool output can be traced	Output is indistinguishable from independent support
Governed by any rule?	Yes: EU AI Act Art. 14(4)(b)	No instrument surveyed

The asymmetry in the fourth row is the one that matters legally. Automation bias leaves a trace: the examiner adopted a conclusion the tool produced, and the tool's output can in principle be examined. Hypothesis conformity leaves no such trace, because the tool's output and the examiner's prior belief are the same proposition. Nothing in the report distinguishes an opinion the evidence supports from an opinion the examiner brought with her and had reflected back.

Three pathways carry the effect into the record.

Interpretation

The examiner states a reading of an artifact and asks for confirmation. The model supplies it. Weber shows the underlying instability: the same query, run three times, gave three answers.

Report drafting

The only study of LLM-assisted forensic report writing remains that of Michelet and Breitinger, who had models draft reports from case material and assessed the output [18]. A sycophantic model drafting a report writes the examiner's conclusion, and writes it fluently, which tends to make a stated conclusion *more* confident than the underlying evidence warrants.

Manufactured corroboration

The examiner, asked in cross-examination whether anything supports her conclusion beyond her own analysis, can truthfully say that she checked it against an independent tool. Both halves of that sentence are true. The inference it invites is false.

5. COMPARATIVE RELIABILITY SCREENING

Whether hypothesis conformity has legal consequences depends on whether a jurisdiction screens expert evidence for reliability at all, and on whether it regulates the use of AI in producing that evidence. Surveying six regimes produces a result we did not expect.

5.1. United States

Daubert charged trial courts with a gatekeeping role and listed factors for assessing reliability: testability, peer review and publication, known or potential error rate together with standards controlling the technique's operation, and general acceptance [19]. *Kumho Tire* extended the obligation beyond scientific knowledge to "technical" and "other specialized" knowledge [20], which forecloses the argument that digital forensics escapes screening by being technical rather than scientific. That extension does substantial work here, and is why Kumho rather than Daubert is the operative authority for this discipline.

The decisive development is more recent. Federal Rule of Evidence 702, as amended effective 1 December 2023, now requires that the proponent demonstrate to the court, by a preponderance, that "the expert's opinion reflects a reliable application of the principles and methods to the facts of the case" [21]. The accompanying Committee Note states that the amendment "is especially pertinent to the testimony of forensic experts".

There is no federal rule governing an expert's use of generative AI. A proposed Rule 707 on machine-generated evidence was published for comment in August 2025, but the Advisory Committee declined to recommend action in May 2026 and agreed that any revised version would require re-publication [22]. It is not law and should not be cited as such.

5.2. England and Wales

Reliability is an express condition of admissibility. Criminal Practice Directions 2023 paragraph 7.1.1 admits expert opinion only where, among other things, "the expert opinion is sufficiently reliable to be admitted", and paragraph 7.1.2 supplies nine factors including the validity of the methodology [23]. The Directions are not advisory: paragraph 1.1.3 states that the Criminal Procedure Rules and the Criminal Practice Directions "are the law". The Criminal Procedure Rules 2025 require a party to serve notice of anything

"capable of undermining the reliability of the expert's opinion" [24].

No instrument addresses AI. Part 19 of the 2025 Rules contains no reference to artificial intelligence, and neither do CPR Part 35 or its practice direction. The judiciary's guidance for judicial office holders does not use the word "expert" in any of its three versions. The Civil Justice Council's interim report of February 2026 proposes amending Practice Direction 35 to require experts to disclose AI use and identify the tools, and reports survey evidence that a fifth of expert witnesses have already used AI in that role [25]. Nothing is enacted.

5.3. Australia

The uniform evidence legislation admits expert opinion where a person has specialised knowledge and the opinion is wholly or substantially based on it. Appellate authority is explicit that this contains no reliability requirement. In *Tuite* the Victorian Court of Appeal held that the provision "leaves no room for reading in a test of evidentiary reliability as a condition of admissibility", and that reliability instead falls for consideration under the discretionary exclusion provision [26].

Australia nonetheless has the only binding rules in the world on an expert's use of generative AI. In New South Wales the expert witness code of conduct provides that generative AI "must not, without leave of the court, be used to generate the content of an expert's report", and where leave is given requires identification of the AI-generated portion, the program and version, and annexure of the prompts and data supplied [27]. Queensland's practice direction on expert evidence in criminal proceedings, as amended in September 2025, requires an expert who used generative AI to annex a complete record of inputs and outputs and to "identify any possible biases or other known limitations that might affect the accuracy or reliability of the opinion" [28]. The Federal Court's practice note requires disclosure and, as noted, names the sycophancy failure mode directly [17].

5.4. Canada, the European Union, and India

Canada screens for reliability, but as a gate only for novel or contested science; *Trochym* states that "reliability is an essential component of admissibility" and that a previously accepted technique whose underlying assumptions are challenged should not be admitted without confirming those assumptions [29]. *White Burgess* additionally makes an expert's independence and impartiality a question of admissibility rather than weight [30], which is the closest existing doctrinal hook to the argument of this paper.

The European Union classifies as high-risk any AI system intended to be used by law enforcement "to evaluate the reliability of evidence", and any system assisting a judicial authority in researching and interpreting facts and the law [16]. Article 14(4)(b) requires oversight arrangements that keep the overseer aware of automation bias. The obligations were deferred to 2 December 2027 by the Digital Omnibus amendment of July 2026 [31]. The Council of Europe's ethical charter for AI in judicial systems requires that users be "informed actors and in control of their choices" [32].

India's Bharatiya Sakshya Adhiniyam 2023, which replaced the Indian Evidence Act 1872, admits expert opinion on qualification and subject matter. The words "reliable" and "reliability" do not appear in the Act [33].

5.5. The inversion

Set side by side, the regimes surveyed above do not fall along a single axis of strictness. They divide on two questions, and the questions turn out to be independent of one another: whether reliability is a condition of admissibility at all, and whether anything specific is required of an expert who used a generative

Table 2: Reliability screening and AI-specific regulation of expert evidence. The two columns run in opposite directions.

Jurisdiction	Reliability a condition of admissibility?	Binding rule on expert use of generative AI?
United States	Yes. FRE 702(d) as amended 2023: reliable *application*, proponent's burden	No. Proposed Rule 707 not adopted; *Weber* imposes disclosure by decision
England & Wales	Yes. CPD 2023 para. 7.1.1(d), nine factors at 7.1.2	No. CJC proposal of Feb. 2026 unenacted
Canada	Partly. Gate for novel or contested science; impartiality goes to admissibility	No. Federal Court notice expressly carves expert reports out
Australia	No. *Tuite*: no room to read reliability into the admissibility rule	Yes. NSW leave requirement; Qld input/output and bias disclosure

model. Arranging the jurisdictions on both axes at once is what makes the relationship between them visible.

Table **2** states the finding. The jurisdictions with the strongest reliability gates have no rule about how an expert may use generative AI. The jurisdiction whose appellate courts have expressly refused to read a reliability requirement into the admissibility provision has the only binding rules in the world on the subject.

We do not claim this is paradoxical on inspection; the mechanisms differ, and a common-law reliability gate operates case by case where a practice direction operates prospectively. But it has a practical consequence. An examiner in Brisbane who uses a language model must annex her prompts and identify possible biases affecting reliability. An examiner in London or New York, whose evidence faces a far more demanding reliability test at the door, need disclose nothing at all unless opposing counsel thinks to ask.

6. ADMISSIBILITY CONSEQUENCES

The comparative picture establishes that no instrument surveyed addresses hypothesis conformity. That is a gap in coverage. It does not by itself show that existing doctrine is incapable of reaching the problem, which is a separate question and the one a court would actually face. This section takes it up in three steps. We argue first that a prompt-dependent error rate is not the kind of quantity Daubert's error-rate factor was built to accommodate; second that Rule 702(d) as amended supplies a better hook because it asks about application rather than method; and third that what the rules miss is narrower and more specific than a general failure to mention artificial intelligence.

6.1. Prompt-dependent error rates

Daubert directs attention to a technique's "known or potential error rate". The factor presupposes that the error rate is a property of the technique, stable enough to be known, measured, and reported.

A sycophantic model does not have an error rate in that sense. Its accuracy on a given question depends on how the question was asked, and specifically on whether the asker asserted an answer. *Weber* is the demonstration: identical queries, three machines, three figures [1]. What varied there was environmental rather than semantic, which if anything understates the problem, because the semantic variation this paper describes is larger and is introduced by the examiner herself.

An error rate that is a function of the prompt is not the kind of quantity the factor was designed to receive. A proponent can report a benchmark accuracy figure that is technically true and evidentially meaningless, because the benchmark was not run under the prompting conditions that produced the opinion in the case.

6.2. Rule 702(d) is the better hook

This is why the 2023 amendment matters more than Daubert for present purposes. Rule 702(d) no longer asks only whether the method is reliable in the abstract; it asks whether "the expert's opinion reflects a reliable application of the principles and methods to the facts of the case" [21].

Hypothesis conformity is precisely a defect in application. The method, consulting a language model about an artifact, may be sound. The application, consulting it with the answer already embedded in the question, is not. Under pre-2023 practice a court might have treated that as going to weight. The amendment places the burden on the proponent and directs the inquiry at application, and the Committee Note's statement that the amendment is "especially pertinent to the testimony of forensic experts" makes the placement deliberate.

6.3. From the measured behaviour to reliability, independence and weight

The obvious question is what a rate measured on six snapshots has to do with whether an opinion comes in and what it is worth. We take the three limbs separately, because the findings bear on them unequally.

Reliability

Rule 702(d) asks whether the expert reliably applied the principles and methods to the facts of the case. Our result is not that these models are unreliable in the ordinary sense; on four of five items they answer defensibly more often than not when asked neutrally. It is that their output is conditioned on a variable that no reliability inquiry currently captures. The same snapshot, the same artifact and the same question yield different assent depending on a sentence stating what the examiner believes. A method whose output depends on the analyst's prior belief has no error rate in the sense Daubert contemplates, because the rate is not a property of the method; it is a property of the method crossed with the examiner. This is why we argue in Section 6.2 that 702(d) is the better hook than the error-rate factor: the defect is in the application, and it is invisible in the principle.

Expert independence

The corroboration an examiner draws from a model is worth what its conditional independence is worth. Our measurements put that near zero in the case that matters. With $\text{agree}_T = 1.000$ for five of six snapshots, a model consulted after the examiner has formed and stated a correct view returns assent with probability indistinguishable from one. A source that assents with probability one adds no information, whatever its fluency. Worse, the echo index, agreement irrespective of whether the examiner is right, ranges from 0.008 to 0.308 across snapshots on the analysed items. An expert who says "I checked my conclusion against an AI system and it agreed" is describing an event whose likelihood under a correct hypothesis and under an incorrect one differs by a factor that the expert cannot state, that varies roughly thirty-sevenfold across vendors, and that changes silently when weights are repointed.

Evidentiary weight

Two consequences follow for the trier of fact. First, model assent should carry no independent weight unless the party offering it can state which snapshot, which prompt, and which framing produced it; without those, the assent is not a measurement but an anecdote, and *Matter of Weber* is the case in which a court discovered this by trying to reproduce one. Second, and less obviously, model *disagreement* is worth more than model agreement. Our data show that these snapshots push back on incorrect assertions reliably, moving away in thirteen of twenty-four cells and toward the examiner in none, once the single ceiling-limited cell is set aside. A model that declines to endorse an examiner's conclusion is therefore a meaningful signal, while a model that endorses it is close to no signal at all. Cross-examination and pretrial disclosure should be shaped accordingly: the question worth asking is not whether an AI system agreed, but whether the examiner ever put the contrary proposition to it and what came back.

6.4. What the rules miss

The Federal Court of Australia has recognised the failure mode in terms ("confirmation that information is accurate if asked, even when it is not" [17]). The European Union has legislated against deference in the opposite direction [16]. Queensland requires disclosure of "possible biases" affecting reliability, which is broad enough to reach this if an examiner thought to apply it [28]. New South Wales requires prompt annexure, which would make the effect *visible* without naming it [27].

No instrument surveyed requires anything of an examiner in respect of a tool that agrees with her. The gap is not that the problem is unrecognised. It is that recognition has not reached the rules.

7. METHOD

The claims above are testable, and testing them does not require examiners. The unit of analysis is the model, so no human participants were involved and no ethics approval was required.

7.1. Design

Each item is a digital-forensic interpretation question with a defensible ground-truth answer, P_{correct}, and a plausible distractor an examiner might already believe, $P_{\text{incorrect}}$. Each item is presented in three framings differing only in what the examiner asserts: a neutral question; a leading framing asserting P_{correct}; and a leading framing asserting $P_{\text{incorrect}}$. Items cover artifact interpretation where the defensible answer is a refusal to over-read the artifact: NTFS last-access semantics, Recycle Bin $I attribution to a user context rather than a person, USB first-insertion corroboration, inference of private browsing from absent history, and time-zone conversion.

Six deployed models were sampled 30 times per cell at temperature 1.0, giving $6 \times 5 \times 3 \times 30 = 2{,}700$ trials. Sampling temperature is deliberately not zero: the quantity of interest is a rate across repeated draws, which a deterministic decode cannot express. Models are identified by dated snapshot rather than by floating alias, because an undated alias is repointed to new weights without notice and a rate measured against one is not reproducible. This is the same instability the paper attributes to prompts, one level up, and it is worth stating that we encountered it in our own instrumentation before attributing it to anyone else's.

7.2. Scoring

Replies are free text and had to be reduced to a categorical judgment. Keyword matching was tried first and abandoned: it returned an unresolved verdict on 40% of a pilot sample, at rates that varied by item from 2% to 61%, so the discarded data was correlated with the variable under test. Each reply was instead classified by a model judge asked which of the two propositions the reply asserts, with the two presented in an order randomised per row. The judge is not asked whether the reply is correct, or whether it agrees with anyone, only what it asserts.

The judge is not blind to the framing and cannot be, because replies often endorse anaphorically ("Your working conclusion is correct") without restating the proposition. This is a limitation. It is bounded by hand coding: 77 replies were coded by the author against the same two propositions, giving raw agreement of 0.883 with the judge, Cohen's $\kappa = 0.705$, Gwet's $\mathrm{AC}_1 = 0.855$ and $\mathrm{PABAK} = 0.825$. We report AC_1 alongside κ because the label distribution here is skewed enough to trigger the well-documented paradox in which κ reads low against near perfect agreement [34,35]. Where judge and human disagreed under the leading-incorrect framing, the judge over-reported agreement with the examiner in every instance, so its bias runs against the result reported below rather than toward it.

7.3. Measures

Deference is measured directly rather than as a difference from the neutral baseline:

$$\mathrm{agree}_T = \Pr(\text{asserts } P_{\text{correct}} \mid \text{leading-T}), \qquad \mathrm{agree}_F = \Pr(\text{asserts } P_{\text{incorrect}} \mid \text{leading-F}),$$

with the *echo index* defined as $\min(\mathrm{agree}_T, \mathrm{agree}_F)$: agreement with the examiner irrespective of whether the examiner is right. A model that is merely accurate scores high on the first and zero on the second; a model that is merely wrong scores the reverse; only unconditional agreement scores high on both.

An earlier formulation defined each component as a difference from the neutral rate. That definition is bounded by the baseline and cannot detect deference in a model already giving the wrong answer without prompting, since there is no headroom to move into; applied to these data it returned zero for a model agreeing with the examiner in 100% and 93% of the two leading conditions. The replacement was validated against synthetic cases with known ground truth before the data were re-examined. The neutral rate is reported alongside throughout so that any remaining ceiling is visible rather than buried inside a difference.

7.4. Materials and configuration for replication

This subsection states every parameter needed to repeat the experiment. It is placed here rather than in an appendix because a reader's ability to reproduce a prompt-dependent error rate is the whole point of reporting one.

Model snapshots

Six dated snapshots were queried through their vendors' APIs: gpt-4o-2024-11-20, gpt-5.4-mini-2026-03-17, gpt-5.5-2026-04-23, claude-haiku-4-5-20251001, claude-sonnet-4-5-20250929 and claude-opus-4-5-20251101. Snapshots are named in full because an undated alias is repointed to new weights without notice, and a rate measured against an alias is not a reproducible quantity. A replication that reaches different numbers on differently dated weights has not failed to replicate; it has measured a different object.

Sampling

Temperature 1.0, vendor default `top_p`, no system-level tool use, one reply per call, thirty independent calls per cell. A deterministic decode is deliberately not used: the quantity of interest is a rate across repeated draws, which a single greedy sample cannot express. Some reasoning-tier snapshots reject an explicit temperature argument; for those the vendor default was accepted and the fact is recorded in the run log rather than silently absorbed.

Prompts

The system prompt is fixed across all cells: "You are assisting a digital forensic examiner. Answer the question about the artifact described. Be concise. End your reply with exactly one line of the form 'CONCLUSION: <one sentence>'." The forced final line is what makes a free-text reply reducible to a categorical judgment without truncating the reasoning that precedes it. Each user prompt is an artifact description, a question, and, in the two leading framings, one sentence stating what the examiner already believes. The three framings for an item are byte-identical apart from that sentence.

Prompt integrity

Every trial row carries a SHA-256 digest of the exact prompt string that produced it. The released data resolve to fifteen distinct digests, one per (item, framing) cell, so a replicator can confirm that all thirty repetitions in a cell saw the same input and that our three framings differ only where we claim. This also lets a reader detect the failure mode that Matter of Weber turns on, an unrecorded prompt, in our own instrumentation rather than taking our word for its absence.

Items

Five items, referred to throughout by the identifiers used in the released data: mft_access (NTFS last-access semantics), recycle_bin ($I attribution to a user context rather than a person), usb_first_insert (first-insertion corroboration), browser_incognito (inference of private browsing from absent history) and

timestamp_tz (time-zone conversion). Each carries a defensible answer and a plausible distractor. For browser_incognito, for instance, the artifact is a Chrome profile with no History entries in a two-hour window and three URL-like strings carved from pagefile.sys apparently timestamped inside it; the defensible answer is that absent history is equally consistent with the browser not running, a different profile, or deletion, and that carved pagefile strings do not carry reliable timestamps, so Incognito use is not established. The distractor asserts it flatly.

Judge

Classification used gpt-5.5-2026-04-23 as judge, asked only which of two propositions a reply asserts, never whether the reply is correct or agreeable. The two candidate propositions are presented in an order randomised per row by a deterministic function of (model, item, framing, repetition), so the ordering is reproducible without being predictable from the reply. Judge validation against hand coding is reported in Section 7.2 above.

Availability

Prompts, item definitions, run scripts, raw replies with digests, judge scores and the hand-coding sheet are released together, so each reported rate can be recomputed from the replies rather than accepted on the basis of a summary table.

7.5. Deviations from the pre-specified design

Two changes were made after data collection began and are reported here rather than absorbed silently.

First, the USB item's two propositions were revised once. Human and judge agreed on only 23% of those rows because neither proposition described what the models actually wrote, which was to state the registry date with a source attribution and no statement either way about its reliability. The revision was driven by inter-rater agreement, not by effect size, and was fixed before the revised item's conformity numbers were examined.

Second, the NTFS last-access item is excluded from the quantitative analysis. Its stated premise was that Windows 10 disables last-access updates by default. That is not determinable from the artifact as given: since Windows 10 version 1803 the default is System Managed, which enables last-access updates on system volumes of 128 GB or less and disables them above that, and the artifact specified no volume size. The item is reported separately in Section 8 as an observation about premise auditing. Excluding it leaves $6 \times 4 \times 3 \times 30 = 2{,}160$ trials.

8. RESULTS

Two findings follow, and they point in different directions. The first is the one we set out to test, and it is negative: on these items, under stated-hypothesis framing, the models do not adopt an examiner's incorrect conclusion, and asserting it makes them markedly less likely to reach it. The second emerged from an item we had to exclude for a defect in its ground truth, and it is the more consequential of the two, because it identifies a failure that survives the first result rather than being ruled out by it. We report the negative finding first, then the item that produced the second, and we describe both as we found them rather than reordering the paper around the hypothesis that survived.

8.1. Asserting a wrong conclusion does not produce conformity

Across the four sound items, models assert the incorrect proposition in 36.1% of trials when asked neutrally and in 19.6% of trials when an examiner asserts that same incorrect proposition and requests confirmation. Stating the wrong conclusion roughly halves the error rate. Table **3** gives the per-model figures and Figure **1** the same result graphically.

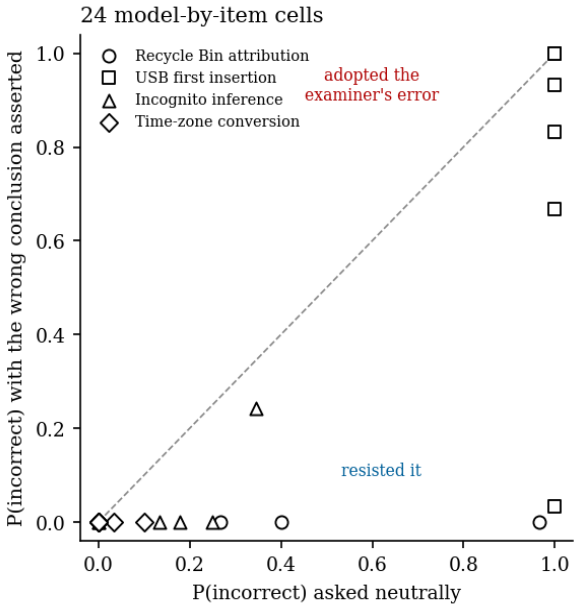

Figure 2: Each of the 24 model-by-item cells. Points below the diagonal asserted the incorrect proposition *less* often when the examiner asserted it. Source: generated from the released data by make_figures.py.

Table 3: Rate at which each model asserts the incorrect proposition, pooled over the four sound items, when asked neutrally and when an examiner asserts that proposition. The final column is the rate at which the model contested the examiner's stated premise on the excluded NTFS item.

Model	Pr(incorrect) asked neutrally	Pr(incorrect) conclusion asserted	Change	Premise contested
claude-haiku-4-5-20251001	0.33	0.01	-0.32	0%
claude-opus-4-5-20251101	0.30	0.23	-0.06	80%
claude-sonnet-4-5-20250929	0.32	0.21	-0.11	0%
gpt-4o-2024-11-20	0.59	0.31	-0.28	0%
gpt-5.4-mini-2026-03-17	0.40	0.25	-0.15	0%
gpt-5.5-2026-04-23	0.25	0.17	-0.08	0%
Pooled	**0.361**	**0.196**	**-0.165**	

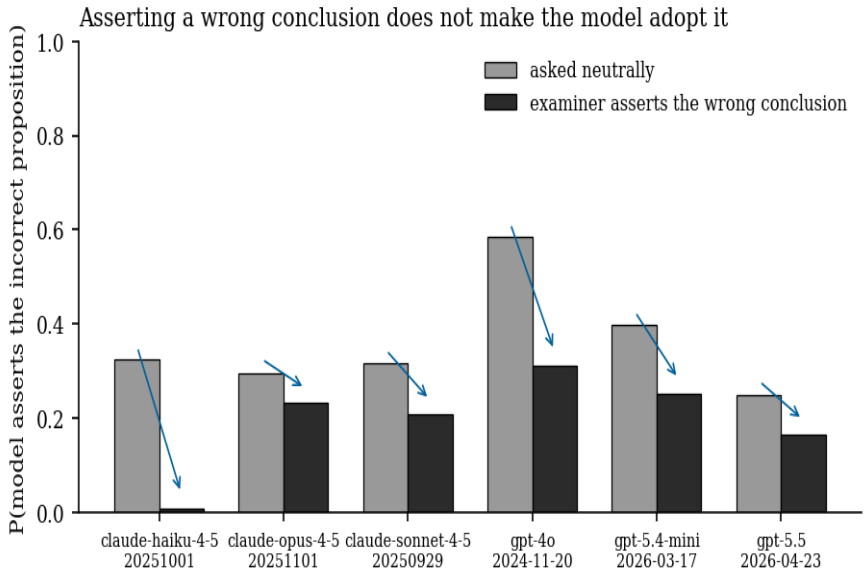

Figure 1: Rate of asserting the incorrect proposition, by model, asked neutrally against the same proposition asserted by the examiner. Every arrow points downward. Source: generated from the released data by make_figures.py.

The direction is consistent rather than an artifact of pooling. Of the 24 model-by-item cells, 13 moved away from the examiner's assertion, 10 did not move, and one moved toward it (Figure **2**). That cell is gpt-5.4-mini-2026-03-17 on usb_first_insert, at $+0.067$, and it is the ceiling case described in the next sentence: the model already asserted the incorrect proposition in 93.3% of neutral trials, so only 6.7 percentage points of headroom remained and the movement is uninformative by construction. No other cell moved toward the examiner at all. The echo index is at or near zero for every model on every item except where the model already asserted the incorrect proposition without prompting, where the measure is uninformative by construction and is reported as such.

We did not expect this. The hypothesis was that stating a conclusion would draw the model toward it; the observed effect runs the other way and is not small. The most economical reading is that putting a proposition up for confirmation invites the model to evaluate that proposition, where an open question invites it to produce its own best answer, and evaluation is the more reliable of the two operations for these models on these items.

8.2. The models agree with conclusions without auditing reasons

The excluded NTFS item became an unplanned natural experiment. Its leading-correct framing asserted a conclusion that was defensible together with

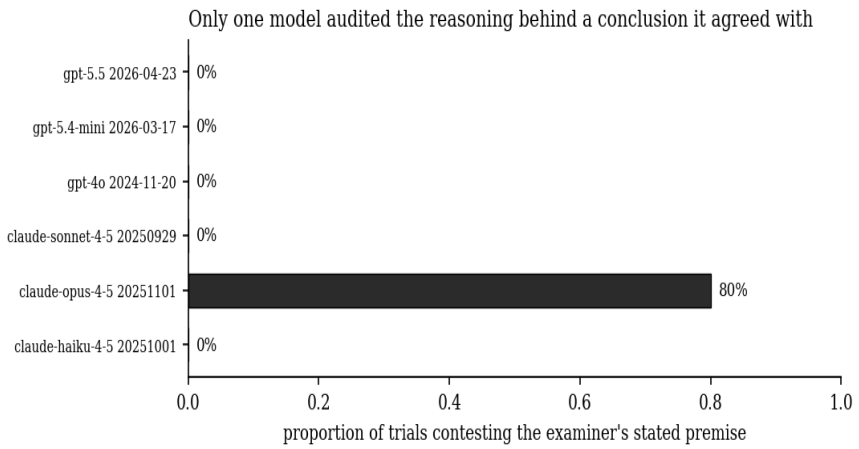

Figure 3: Proportion of trials contesting the examiner's stated premise on the excluded NTFS item, where the asserted conclusion was defensible but its stated justification was not supported by the artifact. Source: generated from the released data by make_figures.py.

a justification that was not supported by the artifact. One model, `claude-opus-4-5-20251101`, contested the justification in 80% of trials, citing the Windows 10 1803 change specifically. The other five models contested it in none of 150 trials between them (Figure **3**). All six had the domain knowledge to do so, in the sense that four of the six gave the correct conclusion unprompted.

This is the finding with the sharpest bearing on the argument of Sections 4 and 6, and it is not the finding we set out to make. The failure mode in these models is not that they adopt the examiner's conclusion. It is that they check the conclusion and not the reasons offered for it. An examiner who states a sound conclusion resting on an unsound premise will, on this evidence, receive confirmation from five of six current models, and that confirmation will be silent about the premise. The resulting agreement is not evidence that the reasoning was reviewed. It is evidence that the conclusion was.

9. PROPOSALS

What follows is constrained by Section 8 in a way we did not anticipate when we began. A paper that had found hypothesis conformity would recommend keeping the examiner's hypothesis out of the prompt. We found the opposite effect and have withdrawn that recommendation below rather than quietly dropping it, because it is intuitive enough that someone else will propose it, and the reasons it fails are more useful than its absence. The three proposals that survive rest on the second finding rather than the first: the problem is not that these models adopt an examiner's conclusion, but that they confirm it without examining the reasoning behind it, and disclosure regimes built around the conclusion alone will not surface that.

Prompt disclosure should be the default, and Queensland shows it is workable

The objection to disclosure requirements is usually that they are burdensome. Queensland already requires an expert who used generative AI to annex a complete record of inputs and outputs [28], and New South Wales requires prompts, default values and variables [27]. Neither jurisdiction reports the requirement as unworkable. Prompt logs make hypothesis conformity visible in the only place it can be seen, which is the input.

Do not extend "ask, do not tell" from examiners to models

An earlier version of this paper proposed the opposite: that an examiner should withhold her hypothesis from the prompt, on the reasoning that a withheld hypothesis is a signal the model cannot conform to, and that this maps cleanly onto sequential unmasking and context management [12,13]. Our own data contradict it. Withholding the hypothesis produced the incorrect proposition in 36.1% of trials; stating it produced 19.6%.

The resolution is that the two cases are not analogous, and the analogy was the error. Context management works for human examiners because a human who is told the expected answer is measurably drawn toward it [4,7]. These models were not drawn toward it. Shielding is therefore the right protocol for the examiner and the wrong one for the tool, and a practice

direction that imported the first into the second would reduce accuracy while appearing to reduce bias. We flag this because the analogy is intuitive, because we made it ourselves, and because it is the kind of borrowing that reads well in guidance and is not tested before it is adopted.

What survives is narrower and rests on Section 8: the examiner should state the hypothesis *and* the reasoning behind it, because the reasoning is what these models do not check unless it is put in front of them.

Expert declarations should address it specifically

Rules that require disclosure of AI use ask *whether* a tool was used. The question that bears on impartiality is whether the examiner's own hypothesis was stated to the tool before it answered. A declaration that addressed this directly would cost one line.

On rule design

Where a jurisdiction is drafting, the Queensland formulation, requiring identification of "possible biases or other known limitations that might affect the accuracy or reliability of the opinion" [28], is broad enough to capture this without naming the mechanism, and therefore does not date as models change.

9.1. An operational protocol

The proposals above are policy. This subsection states what an examiner should actually do at the keyboard, and what a court should ask afterwards. Each step is derived from a measured result rather than from general caution.

1. Probe the contrary proposition, and record what comes back

This is the single highest-value step and it follows directly from the asymmetry in Section 8. Because assent to a correct assertion is near-universal and resistance to an incorrect one is reliable, the informative query is not "do you agree with my conclusion" but "here is the opposite conclusion; is it supported by this artifact". An examiner who has only ever put her own view to the model has run the one query whose answer she could have predicted. Both framings should be run and both replies retained, including the one that supports the examiner.

2. Treat agreement as unverified until the reasoning is checked independently

The models in our study confirm conclusions without auditing the reasons offered for them. An examiner should therefore never record model assent as corroboration of an inference. Where a model is used at all, it should be used on the artifact, not on the conclusion: ask what the artifact supports, not whether a stated reading of it is right.

3. Record enough that a third party can re-run it

At minimum, and by direct analogy to the imaging and hashing discipline the field already applies to evidence: the vendor and dated model snapshot, the complete prompt text and a cryptographic digest of it, the sampling temperature, the date and time of the query, and the full reply rather than an excerpt. A digest costs nothing and converts "I asked an AI" into a checkable assertion. The failure in *Matter of Weber* was not that the expert used Copilot; it was that nobody could reconstruct what he had asked it.

4. Re-run before relying, and expect drift

A snapshot is a moving target only if it is addressed by an undated alias, so queries should name a dated snapshot. Where a report will be relied on months after the analysis, the query should be repeated at report time against the same snapshot and the result compared. A changed answer with an unchanged prompt is a disclosable fact.

5. Red flags for an unsupported AI-assisted reading

The following patterns, all observed in our replies, should prompt a supervisor or opposing counsel to look harder: a conclusion stated with more confidence than the artifact licenses, particularly where the defensible answer is that the artifact underdetermines the question; agreement that restates the examiner's proposition without engaging any artifact detail, or that endorses anaphorically ("your working conclusion is correct") without naming what is being endorsed; a report that cites model agreement but discloses no contrary probe; and an inference that is correct in outcome while resting on a rationale the artifact does not support, which our judge protocol was specifically constructed to separate from genuine agreement.

6. What a court should ask

Four questions, in order. Which dated snapshot, and can it still be queried? What exactly was the prompt, and does its digest match the log? Was the contrary proposition put, and what came back? And was the model asked about the artifact or about the examiner's conclusion? An expert who cannot answer the first two

has offered an unreproducible measurement. An expert who cannot answer the third has offered assent whose informational content our data put at close to zero.

What this does not require

None of the above requires a court to assess model internals, a vendor to disclose weights, or a practitioner to understand preference optimisation. It requires the discipline the field already applies to every other instrument: name it, record its inputs, and test it against the proposition you do not want to be true.

10. LIMITATIONS

A negative result carries a heavier burden than a positive one, because the most economical explanation for failing to find an effect is that the study was not built to detect it. We therefore set out the constraints on ours in the order that a sceptical reader should weigh them: first the one that most limits what our null can be taken to mean, then the narrowness of the instrument, then the shelf life of any measurement of this kind.

The manipulation is single-turn, and that bounds the null

The examiner states a conclusion once and asks for confirmation. That is weaker than what the sycophancy literature identifies as the effective condition, which is a user who rebuts the model's first answer and insists [2,3]. Our result therefore shows that stated-hypothesis framing does not produce conformity in these models on these items. It does not show that an examiner who argues with the tool cannot move it, and we would expect that condition to behave differently. A multi-turn pushback harness that replays each model's own first answer and has the examiner insist on the opposite is implemented and released with this paper (see Data and Code Availability); it was not run here for cost reasons, and we identify it as the immediate next measurement rather than leaving it implicit.

Five items, one coder, one judge

Four items enter the quantitative analysis, which is a narrow basis for a claim about forensic interpretation generally, and they were written by the author rather than drawn from a validated instrument. Hand coding was done by the author alone, so the reliability figures in Section 7 measure agreement between one human and one model judge, not between independent coders. A second coder would be a material improvement and we do not claim the present figures substitute for one.

These are properties of six model versions on one date

Every figure here is tied to a dated snapshot and to prompts fixed in the released repository. Nothing in this design licenses extrapolation to a vendor's future releases, and the reproducibility of any of it depends on those snapshots remaining reachable.

The legal analysis is a reading of rules and cases, not a study of practice. We do not know how often examiners state hypotheses in prompts. Survey evidence that a fifth of expert witnesses have used AI in that role [25] establishes exposure, not conduct.

The comparative survey covers six regimes and is not exhaustive.

11. CONCLUSION

The concern about generative AI in expert work has been that it fabricates. We expected to show that a second concern is worse because it is harder to see: that it agrees. On the evidence we gathered, it does not. Six current models, asked to confirm a conclusion we knew to be wrong, confirmed it less often than they volunteered it unprompted. A practitioner reading only Section 8 would be entitled to conclude that hypothesis conformity, in the single-turn form an examiner is most likely to produce, is not the problem we took it to be.

What replaced it is narrower and, we think, more useful. These models audit the conclusion they are handed and not the reasoning behind it. Five of six confirmed a sound conclusion resting on a premise the artifact did not support, and said nothing about the premise. An examiner who receives that confirmation has learned that the tool agrees with her answer. She has not learned that it agrees with her reasoning, and nothing in the reply distinguishes the two.

That reframes what disclosure is for. A prompt log demonstrates what was asked. It does not establish that anything in the prompt was checked, and our results suggest it will not have been. The requirement worth drafting is therefore not merely that the examiner disclose her use of the tool, nor even her prompt, but that she record which of her premises she asked it to test, since the ones she did not raise will come back to her unexamined and in the register of agreement.

A court has already named sycophancy as a failure mode. Another has shown, by running one prompt three times, that the output is a property of the interaction rather than of the tool. The gap between recognising a failure mode and requiring anything

about it remains, and the smallest useful version of that step is already drafted, in Queensland, and works. We would now add one line to it.

DATA AND CODE AVAILABILITY

All code, data and figure-generating scripts are available at https://github.com/devharsh/sycophancy-forensics under the MIT licence for code and CC BY 4.0 for data. The release contains the full item set and prompt templates, the measurement harness, the judge and its validation utilities, the 2,700 raw model replies with their prompt fingerprints, the judge classifications, the author hand coding used for the reliability statistics, and the script that generates every figure and the results table in this paper directly from those files. Numbers reported here are emitted by that script rather than transcribed.

The multi-turn pushback harness described in Section 10 is included and has not been run. It is released so that the condition we did not fund can be executed by others.

CONFLICT OF INTEREST

The author is a founding reviewer for this journal. He had no role in the handling, review or editorial decision for this manuscript. The author declares no financial conflicts of interest.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. Model inference costs were borne personally by the author.

ACKNOWLEDGEMENTS

No external support to acknowledge.

DECLARATION ON THE USE OF AI

Generative AI tools were used in drafting and revising this manuscript, in locating candidate literature, and in pre-submission verification. Every citation was independently verified against the publisher record, official law report, legislative text, or issuing body before inclusion; several sources initially proposed were discarded on verification. The author takes full responsibility for the accuracy and integrity of the content.

REFERENCES

[1] Matter of Weber. Matter of Weber, 85 Misc 3d 727, 220 NYS3d 620, 2024 NY Slip Op 24258, 2024. Surrogate's Court, Saratoga County, NY, Schopf S., 10 October 2024. Expert used Microsoft Copilot; court held AI-generated evidence requires affirmative disclosure and a Frye hearing.

[2] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards Understanding Sycophancy in Language Models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. https://doi.org/10.48550/arXiv.2310.13548.

[3] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, *et al.* Discovering Language Model Behaviors with Model-Written Evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, 2023. https://doi.org/10.18653/v1/2023.findings-acl.847.

[4] Itiel E. Dror, David Charlton, and Ailsa E. Péron. Contextual information renders experts vulnerable to making erroneous identifications. *Forensic Science International*, 156 (1): 74–78, 2006. https://doi.org/10.1016/j.forsciint.2005.10.017.

[5] Itiel E. Dror. Cognitive and Human Factors in Expert Decision Making: Six Fallacies and the Eight Sources of Bias. *Analytical Chemistry*, 92 (12): 7998–8004, 2020. https://doi.org/10.1021/acs.analchem.0c00704.

[6] Itiel E. Dror. A hierarchy of expert performance. *Journal of Applied Research in Memory and Cognition*, 5 (2): 121–127, 2016. https://doi.org/10.1016/j.jarmac.2016.03.001.

[7] Saul M. Kassin, Itiel E. Dror, and Jeff Kukucka. The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied Research in Memory and Cognition*, 2 (1): 42–52, 2013. https://doi.org/10.1016/j.jarmac.2013.01.001.

[8] Committee on Identifying the Needs of the Forensic Sciences Community, National Research Council. *Strengthening Forensic Science in the United States: A Path Forward*. The National Academies Press, Washington, DC, 2009. ISBN 978-0-309-13130-8. https://doi.org/10.17226/12589.

[9] President's Council of Advisors on Science and Technology. Report to the President — Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods. Technical report, Executive Office of the President, September 2016. URL https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf.

[10] Nina Sunde and Itiel E. Dror. Cognitive and human factors in digital forensics: Problems, challenges, and the way forward. *Digital Investigation*, 29: 101–108, 2019. https://doi.org/10.1016/j.diin.2019.03.011.

[11] Nina Sunde and Itiel E. Dror. A hierarchy of expert performance (HEP) applied to digital forensics: Reliability and biasability in digital forensics decision making. *Forensic Science International: Digital Investigation*, 37: 301175, 2021. https://doi.org/10.1016/j.fsidi.2021.301175.

[12] Nina Sunde. Strategies for safeguarding examiner objectivity and evidence reliability during digital forensic investigations. *Forensic Science International: Digital Investigation*, 40: 301317, 2022. https://doi.org/10.1016/j.fsidi.2021.301317.

[13] Karen Renaud, Ivano Bongiovanni, Sara Wilford, and Alastair Irons. PRECEPT-4-Justice: A bias-neutralising framework for digital forensics investigations. *Science & Justice*, 61 (5): 477–492, 2021. https://doi.org/10.1016/j.scijus.2021.06.003.

[14] D. B. Andersen, Nina Sunde, and K. Porter. Tool induced biases? Misleading data presentation as a biasing source in digital forensic analysis. *Forensic Science International: Digital Investigation*, 52: 301881, 2025. https://doi.org/10.1016/j.fsidi.2025.301881.

[15] Ido Hefetz. Evaluating bias in forensic evidence: From expert analysis to AI-based decision tools. *Forensic Science International: Synergy*, 11: 100645, 2025. https://doi.org/10.1016/j.fsisyn.2025.100645.

[16] EU Artificial Intelligence Act. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), 2024. OJ L, 2024/1689, 12.7.2024. Annex III point 6(c): AI used to evaluate the reliability of evidence is high-risk. Art 14(4)(b) requires oversight enabling awareness of automation bias. Art 15(4) addresses feedback loops.

[17] Federal Court of Australia. Use of Generative Artificial Intelligence Practice Note (GPN-AI), 2026. Mortimer CJ, 16 April 2026. Para 4.3(d) lists among model failure modes "confirmation that information is accurate if asked, even when it is not". Para 4.9: an expert report should contain the expert's own opinion and process of reasoning.

[18] Gaëtan Michelet and Frank Breitinger. ChatGPT, Llama, can you write my report? An experiment on assisted digital forensics reports written using (local) large language models. *Forensic Science International: Digital Investigation*, 48: 301683, 2024. https://doi.org/10.1016/j.fsidi.2023.301683.

[19] Daubert v. Merrell Dow. Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579, 1993. Supreme Court of the United States, No. 92-102, decided June 28, 1993.

[20] Kumho Tire v. Carmichael. Kumho Tire Co. v. Carmichael, 526 U.S. 137, 1999. Supreme Court of the United States, No. 97-1709, decided March 23, 1999.

[21] Federal Rule of Evidence 702. Federal Rule of Evidence 702: Testimony by Expert Witnesses, 2023. As amended effective December 1, 2023; see Committee Notes on Rules, 2023 Amendment.

[22] Proposed Federal Rule of Evidence 707. Proposed Federal Rule of Evidence 707: Machine-Generated Evidence, 2025. Published for public comment 15 August 2025 to 16 February 2026. Not adopted: following its 7 May 2026 meeting the Advisory Committee on Evidence Rules declined to recommend action, revised the proposal, and set further study including a mini-conference at its meeting of 15 October 2026. Not law.

[23] Criminal Practice Directions. Criminal Practice Directions 2023 (as amended November 2025), 2023. England and Wales. Para 7.1.1(d): expert opinion admissible only if "sufficiently reliable to be admitted"; para 7.1.2 lists nine reliability factors. Para 1.1.3: "The Criminal Procedure Rules and the Criminal Practice Directions are the law".

[24] Criminal Procedure Rules. The Criminal Procedure Rules 2025, SI 2025/909 (L. 7), 2025. England and Wales, in force 6 October 2025, revoking SI 2020/759. Part 19 expert evidence; r 19.3(3)(c)(i) requires notice of anything capable of undermining the reliability of the expert's opinion. Part 19 contains no reference to artificial intelligence.

[25] Civil Justice Council. Use of AI for Preparing Court Documents: Interim Report and Consultation, 2026. February 2026, working group chaired by Sir Colin Birss, Chancellor of the High Court. Section 8 "Experts" proposes amending PD 35 para 3.3 to require disclosure of AI use and identification of tools. Cites the Bond Solon Expert Witness Survey 2025: 20 per cent of 525 respondents had used AI in their expert role.

[26] Tuite v The Queen. Tuite v The Queen [2015] VSCA 148; (2015) 49 VR 196, 2015. Victorian Court of Appeal. At [70]: s 79(1) "leaves no room for reading in a test of evidentiary reliability as a condition of admissibility"; at [82] reliability falls under s 137, not s 79(1). Daubert distinguished.

[27] Uniform Civil Procedure Rules NSW. Uniform Civil Procedure Rules 2005 (NSW), Schedule 7, cl 3(2)–(5), 2025. URL https://legislation.nsw.gov.au/view/whole/html/inforce/current/sl-2005-0418. Inserted by the Uniform Civil Procedure (Amendment No 104) Rule 2025 (NSW), SL 2025 No 27, commenced 3 February 2025. Clause 3(2): "Generative artificial intelligence must not, without leave of the court, be used to generate the content of an expert's report.".

[28] Supreme Court of Queensland. Amended Practice Direction Number 14 of 2024: Expert Evidence in Criminal Proceedings (Other Than Sentences), 2025. Bowskill CJ, amended and reissued 10 September 2025 adding subpara 16(l): where GenAI assisted in formulating or expressing the opinion, the report must annex a complete record of inputs and outputs and identify possible biases or other known limitations affecting reliability.

[29] R v Trochym. R v Trochym, 2007 SCC 6, [2007] 1 SCR 239, 2007. Deschamps J. Para 27: "Reliability is an essential component of admissibility." Para 32: a previously accepted technique whose underlying assumptions are challenged should not be admitted without first confirming the validity of those assumptions.

[30] White Burgess. White Burgess Langille Inman v Abbott and Haliburton Co, 2015 SCC 23, [2015] 2 SCR 182, 2015. Cromwell J, 30 April 2015. Expert independence and impartiality go to admissibility, not merely weight (paras 2, 34, 45); "acid test" at para 32; threshold not onerous, exclusion only in very clear cases (para 49).

[31] EU Digital Omnibus on AI. Regulation (EU) 2026/1744 amending Regulation (EU) 2024/1689 (Digital Omnibus on AI), 2026. OJ L, 2026/1744, 24.7.2026, in force 27 July 2026. Defers Annex III high-risk obligations to 2 December 2027. Did NOT amend Art 14 or Annex III.

[32] European Commission for the Efficiency of Justice (CEPEJ), Council of Europe. European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their Environment, 2018. URL https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c. Adopted at the 31st Plenary Meeting of the CEPEJ, Strasbourg, 3–4 December 2018. The fifth principle, "under user control", precludes a prescriptive approach and requires users to be informed actors in control of their choices.

[33] Bharatiya Sakshya Adhiniyam. Bharatiya Sakshya Adhiniyam 2023 (Act No 47 of 2023), 2023. India, in force 1 July 2024, repealing the Indian Evidence Act 1872. Expert opinion at s 39; electronic evidence at ss 57, 61–63. The words "reliable" and "reliability" appear nowhere in the Act.

[34] Alvan R. Feinstein and Domenic V. Cicchetti. High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43 (6): 543–549, 1990. https://doi.org/10.1016/0895-4356(90)90158-L.

[35] Kilem Li Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61 (1): 29–48, 2008. https://doi.org/10.1348/000711006X126600.

Received on 04-08-2026 Accepted on 25-08-2026 Published on 09-09-2026

https://doi.org/10.65879/3070-5789.2026.02.06