The AI Disclosure Paradox: How Epistemic Governance Infrastructure Shapes Credibility Judgments of AI-Assisted Professional Knowledge.

Author: Dr Suela Pirushi

Authors' note: Hypotheses and analysis plans were specified prior to data collection and formally registered on the Open Science Framework prior to journal submission (https://osf.io/ksrc2; DOI: 10.17605/OSF.IO/KSRC2). All materials, measures, stimulus descriptions, and supplementary study details are available at the same location.
Correspondence should be addressed to suelapirushi@yahoo.com

Abstract

How do organisations maintain credible professional judgment when knowledge is co-produced by artificial intelligence? Professional credibility in AI-augmented settings depends not on individual competence alone but on the structural conditions under which content is evaluated and governed. This study tests whether epistemic governance infrastructure causally shapes professional credibility judgments. A registered 2×2×2 between-subjects factorial experiment (Study 1; N = 480) with UK-based professionals held content constant while varying provenance visibility, accountability clarity, and AI disclosure framing. Provenance visibility increased credibility by 23% (η^2p = .21), accountability clarity increased challenge willingness by 40% (d = 0.93), and combined governance produced 53% higher credibility. The study reveals an AI Disclosure Paradox: disclosing AI involvement reduced credibility by 16% under weak governance but had no effect under strong governance, suggesting that mandatory disclosure without accompanying governance may harm professional credibility. Governance effects were 2.6 times stronger for high-stakes claims. Senior professionals responded primarily to provenance cues; junior professionals responded primarily to accountability signals, indicating that governance systems should be experience-differentiated. A supplementary experiment (N = 240) confirmed that contestability mechanisms increased challenge behaviour by 45% (d = 0.83; see Supplementary Materials). These findings establish epistemic governance as a structurally inducible organisational capability with implications for AI policy, leadership theory, and organisational training design.

Keywords: epistemic governance, credibility, provenance, accountability, AI disclosure.

Organisational Relevance and Contribution Statement

This research addresses a pressing organisational challenge: maintaining the credibility of professional outputs when AI co-produces knowledge claims. The findings demonstrate that specific, implementable governance mechanisms, source tracking, named accountability, and structured challenge rights, produce large, measurable improvements in professional evaluation of AI-assisted content. The AI Disclosure Paradox has immediate policy relevance: mandatory transparency without governance infrastructure may reduce rather than increase trust. The differential sensitivity finding provides actionable design guidance for governance systems serving professionals at different career stages.

Introduction

Generative artificial intelligence is transforming professional knowledge production. Large language models now draft market analyses, generate regulatory summaries, construct financial narratives, and populate client-facing proposals across professional services sectors (Eloundou, Manning, Mishkin, & Rock, 2023; Dell'Acqua et al., 2023). This transformation creates a governance challenge that existing organisational theories have not addressed: when AI co-produces knowledge claims, traditional epistemic foundations authorship, provenance, accountability, become ambiguous, yet the credibility of professional outputs remains the core product organisations sell (Empson, Muzio, Broschak, & Hinings, 2015).

The challenge is not whether AI-assisted content is accurate but whether professional outputs produced with AI assistance are perceived as credible by the evaluators who must endorse, present, and defend them. Professional credibility is not an intrinsic property of content; it is an evaluative judgment shaped by the epistemic conditions under which content is produced, presented, and governed (Origgi, 2019; Sperber et al., 2010). When those conditions become opaque, as they do when AI generates fluent, authoritative-seeming content without corresponding epistemic foundations, evaluators cannot distinguish well-founded from poorly-founded claims using traditional surface cues (Parasuraman & Manzey, 2010).

Despite growing scholarly attention to AI governance (Floridi et al., 2018; Jobin, Ienca, & Vayena, 2019), research has focused primarily on ethical deployment, bias, fairness, and privacy rather than on the epistemic quality of AI-assisted organisational knowledge. A system can be ethically compliant yet epistemically unsound (Zerilli, Knott, Maclaurin, & Gavaghan, 2019). This gap is consequential: recent qualitative evidence from AI-augmented professional services organisations reveals that firms using identical AI tools experience dramatically different epistemic outcomes depending on the strength of their governance infrastructure.[1]

The present research introduces the concept of epistemic governance infrastructure, the organisational structures through which the credibility of AI-assisted knowledge is established, and tests whether specific mechanisms causally improve professional credibility judgments. Our central claim is that credibility in AI-augmented settings is

structurally inducible: it can be deliberately produced through infrastructure design rather than relying on individual competence, AI model quality, or organisational culture.

Theory and Hypotheses

Credibility as an Organisational Judgment

Our theoretical foundation integrates social epistemology, accountability theory, and research on professional judgment under algorithmic support. Professional evaluators rarely have direct access to the truth value of complex knowledge claims; instead, they rely on epistemic heuristics, signals indicating whether a claim is likely reliable and defensible (Sperber et al., 2010; Origgi, 2019). In human-authored contexts, the author's expertise and professional reputation serve as implicit credibility markers.

AI co-production disrupts these heuristics. Large language models generate fluent, coherent outputs without corresponding epistemic grounding (Bender, Gebru, McMillan-Major, & Shmitchell, 2021; Floridi & Chiriatti, 2020). The surface qualities professionals use to assess credibility, coherence, comprehensiveness, formatting, are precisely the features AI produces regardless of evidentiary quality. This creates an epistemic heuristic gap, increasing reliance on structural governance cues. Governance mechanisms function as credibility scaffolds: they do not guarantee correctness, but increase perceived legitimacy by making epistemic processes visible.

Provenance Visibility

Provenance visibility refers to the degree to which evaluators can reconstruct how a knowledge claim was produced. Social epistemological accounts establish that rational credibility assessment requires information about the conditions of testimony production (Goldberg, 2010; Lackey, 2008), and algorithmic transparency research demonstrates that process visibility increases appropriate reliance (Shin, 2021; Kizilcec, 2016). In AI co-production contexts, provenance visibility restores epistemic traceability that AI generation disrupts. Based on this reasoning, we propose:

Hypothesis 1. AI-assisted judgments accompanied by high provenance visibility will be perceived as more credible than equivalent judgments with low provenance visibility.

Accountability Clarity

Accountability clarity refers to whether responsibility for a knowledge claim is clearly assigned to an identifiable individual. Accountability theory establishes that anticipated accountability increases cognitive effort and self-critical evaluation (Lerner & Tetlock, 1999; Tetlock, 1983). In AI-augmented settings, accountability is frequently diffused across human actors and AI systems (Zerilli et al., 2019). Clear assignment should both increase perceived quality and increase willingness to challenge (Frink & Klimoski, 1998; Hall, Frink, & Buckley, 2017).

Hypothesis 2a. AI-assisted judgments with clear accountability will be perceived as more defensible than judgments with diffuse accountability.

Hypothesis 2b. Clear accountability will increase evaluators' willingness to raise concerns about AI-assisted outputs.

Interaction, Disclosure, and Stakes

Provenance and accountability were theorised to interact: provenance should be more effective when evaluators can direct concerns to a responsible party. AI disclosure framing was theorised to operate through authority attribution: provisional framing should preserve evaluative engagement relative to authoritative framing (Parasuraman & Riley, 1997; Lee & See, 2004). Stakes moderation was expected because governance cues should become more salient under severe consequences.

Hypothesis 3. The positive effects of provenance visibility on credibility will be amplified when accountability is clearly assigned.

Hypothesis 4. AI-assisted judgments framed as provisional inputs will be perceived as more credible than equivalent judgments framed as authoritative outputs.

Hypothesis 5. The effects of governance mechanisms on credibility will be stronger for high-stakes claims.

Taken together, these hypotheses describe a single system in which two governance inputs, provenance visibility and accountability clarity, make independent, additive contributions to the perceived credibility of AI-assisted knowledge; disclosure framing exerts a contingent effect whose direction depends on infrastructure strength; and task

stakes and evaluator experience moderate the strength and routing of these effects. Figure 1 presents this integrative conceptual model.

Method

Design

A 2 (provenance visibility: high vs. low) × 2 (accountability clarity: clear vs. diffuse) × 2 (AI disclosure framing: provisional vs. authoritative) between-subjects factorial experiment with within-subjects stakes moderation. Hypotheses and the analysis plan were specified prior to data collection and formally registered on the Open Science Framework prior to submission (https://osf.io/ksrc2).

Participants

Four hundred and eighty professionals participated, with 60 randomly assigned to each of eight cells. Participants were recruited through professional networks and the Prolific platform, screened for current employment in consulting (n = 230, 48%), legal (n = 149, 31%), or technical advisory (n = 101, 21%) roles with a minimum of six months' experience. Experience ranged from junior (< 3 years, n = 154) through mid-career (3–7 years, n = 197) to senior (7+ years, n = 129). Mean age was 34.2 (SD = 8.1); 52% female. A priori power analysis indicated power > .90 to detect medium effects (f = 0.25) in 2×2×2 ANOVA at α = .05.

Stimulus Materials

All participants reviewed a UK digital health market assessment prepared for a hypothetical client board presentation, developed in consultation with three practising consultants. The assessment contained well-supported claims alongside strategically embedded weakly supported claims, a market sizing estimate from incompatible sources, a regulatory timeline with incorrect dates, a competitor analysis from outdated data, and an unjustified growth projection designed to pass casual review but be identifiable with scrutiny.

In the high-provenance condition, each claim was accompanied by inline annotations showing origin (primary research, AI-generated, analyst synthesis), verification status, and evidentiary foundation. In the low-provenance condition, content was presented

without source information. In the clear-accountability condition, a named individual was identified as epistemic owner with a statement of personal accountability. In the diffuse condition, the document was attributed to a team. AI disclosure was framed as either provisional ("AI-assisted draft, subject to review") or authoritative ("AI-powered analysis").

Manipulation Checks

Manipulation checks asked participants to identify (a) whether source information was present, (b) who was accountable for the document, and (c) how AI involvement was framed. Classification accuracy was 94% for provenance, 91% for accountability, and 89% for AI framing (all χ^2 tests $p < .001$).

Measures

Five dependent variables were measured using scales developed through item generation from qualitative research and pilot testing (N = 40; see https://osf.io/ksrc2 for full items and psychometric details). All exceeded $\alpha > .80$: perceived credibility (4 items, $\alpha = .87$; sample: "The claims in this assessment are supported by adequate evidence"), perceived defensibility (3 items, $\alpha = .84$), challenge comfort (3 items, $\alpha = .82$), endorsement willingness (binary plus 7-point confidence), and epistemic confidence (3 items, $\alpha = .81$). Confirmatory factor analysis established discriminant validity.

Procedure

Participants completed the study online (25–35 minutes). After informed consent, participants read a scenario establishing them as a senior reviewer evaluating an AI-assisted market assessment before client presentation, reviewed the assessment under randomly assigned conditions, evaluated a high-stakes claim (regulatory compliance projection) and a low-stakes claim (market background summary), completed dependent variable measures, answered manipulation checks, and responded to open-ended questions about their evaluative reasoning.

Analytical Strategy

Registered analyses comprised 2×2×2 between-subjects factorial ANOVA for continuous variables, with planned contrasts comparing combined high- versus low-

governance conditions. Within-subjects stakes effects used mixed ANOVA. Binary outcomes used logistic regression. Experience moderation used regression-based approaches (Hayes, 2017). Effect sizes are reported using η^2p, Cohen's d, and odds ratios.

Results

Provenance visibility (H1).

H1 was strongly supported. Provenance visibility produced a large main effect on perceived credibility, $F(1, 476) = 124.3$, $p < .001$, $\eta^2p = .21$. High-provenance participants rated credibility at $M = 5.42$ ($SD = 0.98$) versus $M = 4.41$ ($SD = 1.12$) — a 23% increase. The defensibility effect was similarly large, $F(1, 476) = 98.7$, $p < .001$, $\eta^2p = .17$. Endorsement: 72% endorsed under high provenance versus 51% under low (OR = 2.47, $p < .001$).

Open-response data illuminated the mechanism. High-provenance participants described using source indicators to calibrate evaluative attention, concentrating scrutiny on AI-generated, unverified claims rather than reviewing uniformly. One senior consultant wrote: "The source tags made my review targeted rather than general. I could see immediately which sections needed careful checking." Low-provenance participants reported relying on surface fluency, precisely the features AI produces regardless of evidentiary quality.

Accountability clarity (H2a, H2b).

Both hypotheses were supported with large effects. Voice intent increased by 37 percentage points under clear accountability (71% vs. 34%, OR = 2.13, $p < .001$). Challenge comfort: $F(1, 476) = 89.2$, $p < .001$, $d = 0.93$ (clear: $M = 5.11$; diffuse: $M = 3.64$). Defensibility: $F(1, 476) = 67.4$, $p < .001$, $\eta^2p = .12$. The accountability effect on credibility ($\eta^2p = .09$) was smaller than provenance ($\eta^2p = .21$), indicating independent, non-substitutable contributions.

Open-response data revealed two channels. Clear accountability raised perceived stakes of endorsement and paradoxically increased comfort with challenge by signalling

that scrutiny was structurally expected: "If someone's name is on it, they want me to find problems. That's the whole point."

Interaction effects (H3).

H3 was supported with an additive rather than synergistic pattern. Combined high-provenance and clear-accountability produced credibility of M = 5.84 versus M = 3.82 under combined weak conditions, a 53% improvement. The interaction was significant, $F(1, 476) = 8.42$, $p = .004$, $\eta^2p = .02$, but modest: each mechanism contributed independently and their combination approximately equalled the sum of individual effects, suggesting complementary rather than synergistic governance.

The AI Disclosure Paradox (H4).

H4 revealed the study's most unexpected finding. Rather than a simple framing effect, AI disclosure effects were entirely contingent on infrastructure strength, producing a crossover interaction (see Figure 2). Under weak infrastructure (low provenance, diffuse accountability), disclosing AI involvement reduced credibility by 16%, $F(1, 238) = 14.7$, $p < .001$, $d = 0.52$. Participants reported that disclosure triggered anxiety without evaluative tools: "Knowing AI was involved but having no way to check its work made me trust the document less." Under strong infrastructure, AI disclosure had no significant effect, $F(1, 238) = 0.83$, $p = .36$, $d = 0.08$.

Stakes moderation (H5).

H5 was supported. Governance effects were 2.6 times stronger for high-stakes claims. The stakes × provenance interaction was significant, $F(1, 476) = 22.4$, $p < .001$, $\eta^2p = .05$: provenance effects were $\Delta M = 1.38$ under high stakes versus $\Delta M = 0.53$ under low stakes. The stakes × accountability interaction was also significant, $F(1, 476) = 16.8$, $p < .001$, $\eta^2p = .03$.

Experience moderation (exploratory).

A registered secondary analysis revealed differential sensitivity. Senior professionals (7+ years) were twice as sensitive to provenance cues ($\Delta M = 1.34$ vs. 0.71, interaction $F(2, 474) = 7.83$, $p < .001$). Junior professionals (< 3 years) were 2.3 times more sensitive to accountability cues ($d = 1.21$ vs. 0.53, interaction $F(2, 474) = 9.17$, $p <$

.001). A supplementary experiment (N = 240) further confirmed that contestability mechanisms increased challenge behaviour by 45% (d = 0.83) by reducing perceived social cost rather than increasing analytical capability (see Supplementary Materials for full method and results; also see Figure 3 and Table 1).

General Discussion

This experiment demonstrates that the credibility of AI-assisted professional knowledge is structurally inducible through governance infrastructure design. Provenance visibility (η^2p = .21), accountability clarity (d = 0.93), and their combination (53% credibility improvement) establish that identical AI-assisted content receives dramatically different evaluations depending solely on its governance context.

The Disclosure Paradox and AI Policy

The AI Disclosure Paradox challenges prevailing assumptions in AI governance. Current regulatory trajectories, the EU AI Act, proposed US disclosure requirements, assume that transparency inherently promotes accountability. Our evidence reveals that transparency is beneficial only when accompanied by governance tools enabling informed evaluation. This extends Kizilcec's (2016) work by demonstrating that the transparency–trust relationship is contingent on institutional context. Disclosure mandates should be coupled with governance infrastructure standards; transparency without infrastructure creates what participants described as "anxiety transparency", awareness of AI involvement without the means to evaluate its implications.

Additive Complementarity and Incremental Adoption

The additive interaction pattern suggests that governance dimensions address distinct epistemic vulnerabilities, provenance addresses traceability, accountability addresses ownership, and each produces independent benefits. Organisations can adopt mechanisms incrementally, with each investment yielding independent returns. This lowers the barrier to governance adoption and provides a basis for staged implementation.

Experience-Differentiated Governance

The differential sensitivity finding extends the algorithm aversion literature (Dietvorst, Simmons, & Massey, 2015) by demonstrating that professional responses to AI-assisted content are differentiated by experience in previously unrecognised ways. Senior professionals respond to provenance because they can evaluate evidentiary quality when made visible; junior professionals respond to accountability because they rely on social trust anchors. Optimal governance must be multi-layered, serving evaluators at different career stages.

These findings carry direct implications for how organisations build epistemic capability, and they suggest that uniform AI-governance training is likely to be inefficient. Because senior and junior professionals respond to different cues, capability-building is better targeted than generic. For junior professionals, who anchored their evaluations on accountability signals, the developmental priority is building independent evidentiary judgment so that they are not reliant on knowing whose name is attached to a claim: structured exposure to provenance reasoning, supervised verification of AI-generated material, and explicit instruction in distinguishing surface fluency from evidentiary grounding. For senior professionals, who already read evidentiary quality once provenance is visible, the priority is different, ensuring that provenance information is consistently surfaced in their workflows, since their discrimination is only as good as the cues the infrastructure provides.

This reframes a common assumption in organisational AI adoption. The frequent response of investing in generic prompt-engineering or "AI-literacy" training treats epistemic failure as a uniform skills deficit; the present evidence suggests instead that capability-building should be experience-differentiated and, crucially, paired with infrastructure rather than substituted for it. Training that develops evidentiary judgment has little effect if the workflow never makes provenance visible; conversely, infrastructure that surfaces provenance yields the largest returns when professionals have been trained to act on what it reveals. Capability and infrastructure are complements, not alternatives.

Theoretical Implications

The findings advance theory in three domains. For leadership theory, they demonstrate that leadership influence can be exercised through epistemic infrastructure design in the absence of interpersonal interaction, participants responded to structural features, not leaders.[2] For social epistemology, they provide the first experimental evidence that organisational epistemic infrastructure shapes collective knowledge quality through specific, measurable pathways, operationalising Goldberg's (2010) concept of epistemic dependence networks. For AI governance, the Disclosure Paradox challenges the assumption that transparency inherently promotes accountability, with immediate implications for regulatory design.

Limitations and Future Directions

First, the vignette-based, single-exposure design sacrificed ecological validity for causal precision: it isolates the immediate effect of governance cues but cannot establish whether those effects persist, strengthen, or decay as professionals are repeatedly exposed to AI-assisted content over time, nor can it capture the political costs of challenging senior colleagues' AI-assisted work under sustained organisational pressure. Whether governance sensitivity is durable or itself adapts with exposure is a question a single experiment cannot answer and one that calls for multi-wave field observation.

Second, the sample was drawn entirely from UK-based professionals, which bounds cross-cultural generalisability. Because accountability and challenge norms are culturally patterned, the mechanisms examined here may operate differently across contexts. The accountability and contestability effects in particular may be sensitive to power distance: in higher power-distance settings, named ownership could heighten rather than relieve the social cost of challenging a senior colleague's work, attenuating the voice effects observed here, whereas provenance visibility, which operates on the content rather than the person, may prove more culturally portable. A priority for future research is a multi-site replication across contrasting cultural settings for example, the UK, a higher power-distance Asian context, and the US, to test whether the additive

structure of governance effects, and the Disclosure Paradox in particular, hold cross-culturally or are themselves institutionally contingent.

Third, the study measured credibility perceptions rather than objective evaluative accuracy. Fourth, the within-subjects stakes manipulation, while informative, does not replicate the sustained pressure of real organisational deadlines.

Conclusion

As organisations integrate AI into knowledge-intensive work, the credibility of professional outputs can no longer be assumed, it must be deliberately governed. This experiment demonstrates that specific governance mechanisms substantially improve how professionals evaluate AI-assisted content, and reveals that common transparency practices may harm credibility without accompanying infrastructure. Epistemic governance is a structurally inducible organisational capability, and these findings provide evidence-based design principles for organisations, leaders, and policymakers.

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., ... & Lakhani, K. R. (2023). Navigating the jagged technological frontier. Harvard Business School Working Paper, 24-013.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), 114–126.

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). GPTs are GPTs: An early look at the labor market impact potential of large language models. arXiv preprint arXiv:2303.10130.

Empson, L., Muzio, D., Broschak, J. P., & Hinings, B. (2015). The Oxford handbook of professional service firms. Oxford University Press.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30(4), 681–694.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People — An ethical framework for a good AI society. Minds and Machines, 28(4), 689–707.

Frink, D. D., & Klimoski, R. J. (1998). Toward a theory of accountability in organizations. Research in Personnel and Human Resources Management, 16, 1–51.

Goldberg, S. C. (2010). Relying on others: An essay in epistemology. Oxford University Press.

Hall, A. T., Frink, D. D., & Buckley, M. R. (2017). An accountability account. Journal of Organizational Behavior, 38(2), 204–224.

Hayes, A. F. (2017). Introduction to mediation, moderation, and conditional process analysis (2nd ed.). Guilford Press.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399.

Kizilcec, R. F. (2016). How much information? Effects of transparency on trust in an algorithmic interface. Proceedings of the 2016 CHI Conference, 2390–2395.

Lackey, J. (2008). Learning from words: Testimony as a source of knowledge. Oxford University Press.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50–80.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. Psychological Bulletin, 125(2), 255–275.

Origgi, G. (2019). Reputation: What it is and why it matters. Princeton University Press.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation. Human Factors, 52(3), 381–410.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230–253.

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance. International Journal of Human-Computer Studies, 146, 102551.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. Mind & Language, 25(4), 359–393.

Tetlock, P. E. (1983). Accountability and the perseverance of first impressions. Social Psychology Quarterly, 46(4), 285–292.

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in algorithmic

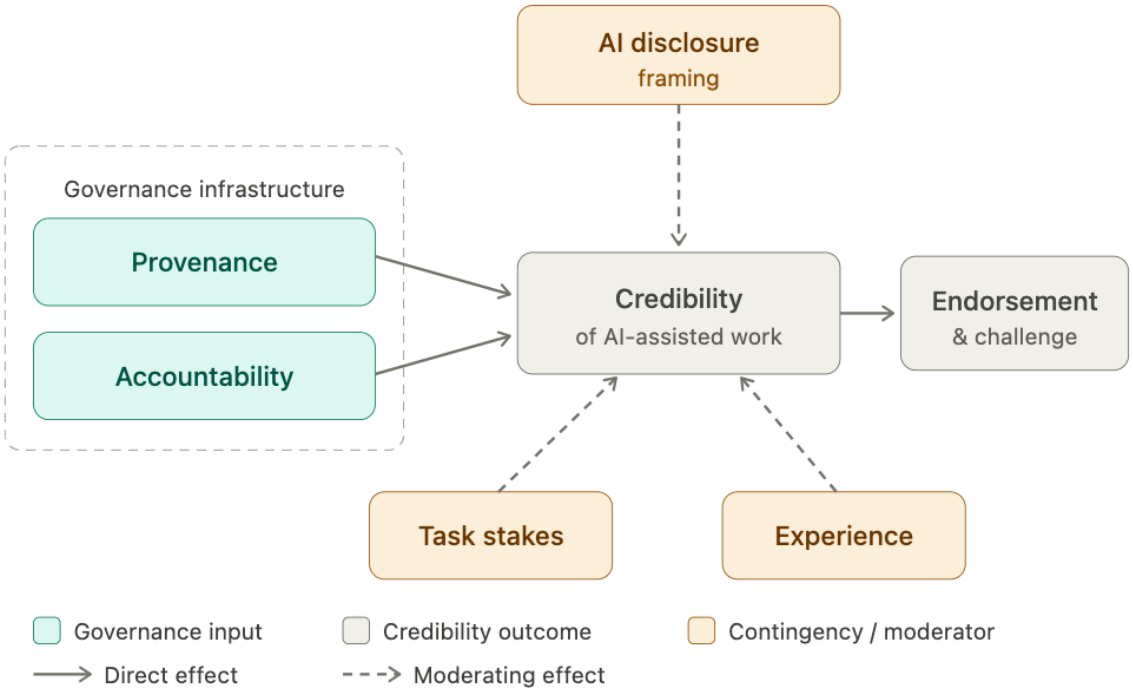
FOOTNOTES

[1.] Related qualitative and longitudinal field studies supporting these experimental findings are reported separately (details available from the author upon request).

[2.] These experimental findings are complemented by a 22-week longitudinal process study conducted across three organisational units, which examines the temporal dynamics of the same governance mechanisms under ecological field conditions and reports multi-wave

Figure 1

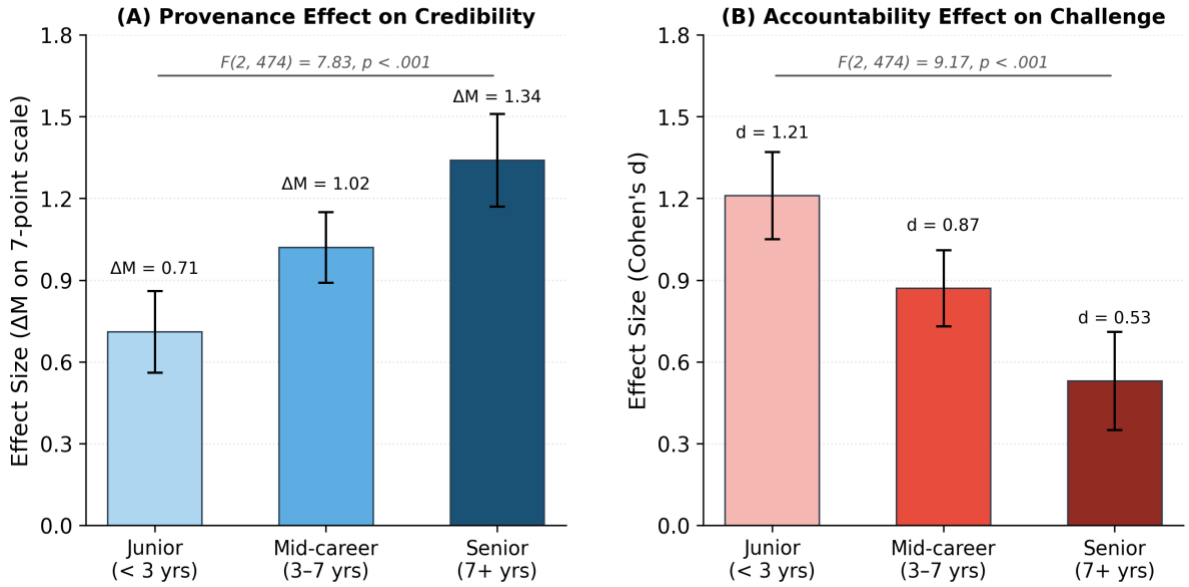
Epistemic Governance Infrastructure: An Integrative Conceptual Model



Note. Provenance visibility and accountability clarity are independent governance inputs that make additive contributions to the perceived credibility of AI-assisted knowledge, which in turn supports endorsement and substantive challenge. AI disclosure framing produces the Disclosure Paradox: a moderating effect (dashed arrow) in which disclosure reduces credibility under weak governance but has no effect under strong governance. Task stakes and professional experience operate as cross-cutting moderators: stakes amplify governance effects 2.6 times, and experience routes which input is most influential (senior professionals respond primarily to provenance cues; junior professionals respond primarily to accountability signals). Solid arrows denote direct effects; dashed arrows denote moderating or contingent effects.

Figure 2

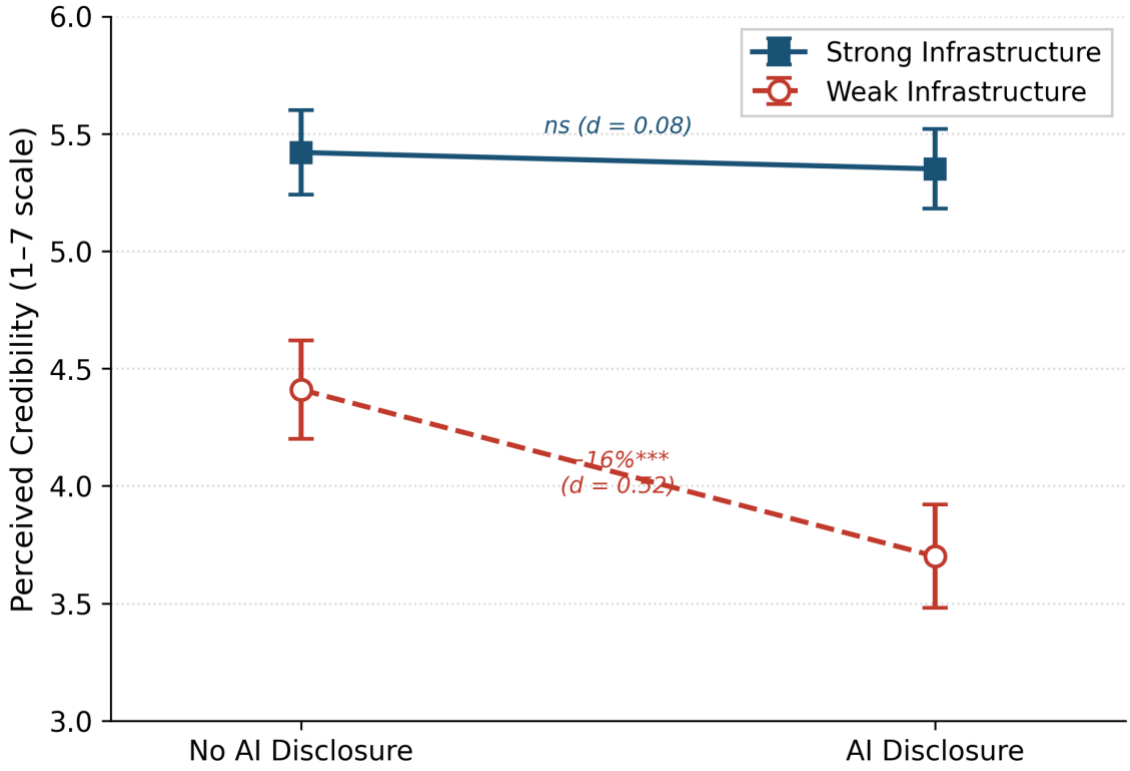
Differential Sensitivity to Governance Mechanisms by Professional Experience



Note. (A) Senior professionals showed 2× sensitivity to provenance cues. (B) Junior professionals showed 2.3× sensitivity to accountability cues. Both interactions $p < .001$. Error bars = 95% CIs.

Figure 3

The AI Disclosure Paradox: Effect of AI Disclosure on Perceived Credibility as a Function of Infrastructure Strength



Note. Strong infrastructure = high provenance + clear accountability. Weak infrastructure = low provenance + diffuse accountability. Under weak infrastructure, AI disclosure reduced credibility by 16% (d = 0.52). Under strong infrastructure, disclosure had no significant effect (d = 0.08). Error bars = 95% CIs. N = 480.

Table 1

Summary of Experimental Effects (Hypotheses 1–5)

Effect	Test Statistic	*p*	Effect Size	Key Metric	Interpretation
H1: Provenance → Credibility	F(1,476) = 124.3	< .001	η^2p = .21	+23%	Large
H1: Provenance → Defensibility	F(1,476) = 98.7	< .001	η^2p = .17	+25%	Large
H2a: Accountability → Defensibility	F(1,476) = 67.4	< .001	η^2p = .12	+21%	Medium-large
H2b: Accountability → Challenge	F(1,476) = 89.2	< .001	d = 0.93	+40%	Large
H2b: Accountability → Voice	OR = 2.13	< .001	37 pp	71% vs 34%	Large
H3: Combined → Credibility	F(1,476) = 8.42	.004	η^2p = .02	+53%	Additive
H4: Disclosure (weak infra)	F(1,238) = 14.7	< .001	d = 0.52	−16%	Paradox
H4: Disclosure (strong infra)	F(1,238) = 0.83	.36	d = 0.08	ns	No effect
H5: Stakes moderation	F(1,476) = 22.4	< .001	η^2p = .05	2.6×	High > Low

Note. N = 480. All tests two-tailed. η^2p = partial eta-squared; d = Cohen's d; OR = odds ratio; pp = percentage points. H4 tested within infrastructure strength conditions. Experience moderation (exploratory): seniors 2× sensitive to provenance, $F(2, 474) = 7.83$, $p < .001$; juniors 2.3× sensitive to accountability, $F(2, 474) = 9.17$, $p < .001$. Supplementary contestability experiment (N = 240): d = 0.83; see Supplementary Materials.