International Journal of Education, Vocational and Social Science
https://e-journal.citakonsulindo.or.id/index.php/IJEVSS
E-ISSN : 2963 4954
IJEVSS

# Communicating Uncertainty in the Age of Generative AI: Effects of Disclosure Strategies on Information Credibility

**Iwan Armawan[1],**
Department of Communication Science, Universitas Mpu Tantular, Jakarta, Indonesia
**Tri Adi Sarwoko[2],**
Department of Communication Science, Universitas Mpu Tantular, Jakarta, Indonesia
**Lasria Sinambela [3],**
Department of Communication Science, Universitas Mpu Tantular, Jakarta, Indonesia
**Dessy Marhandrie[4],**
Department of Communication Science, Universitas Mpu Tantular, Jakarta, Indonesia
**Sephia Wijaya[5]**
Department of Communication Science, Universitas Mpu Tantular, Jakarta, Indonesia

Corresponding Author : iwanaradea84@gmail.com

**Abstract**

*Communicating uncertainty is central to responsible generative artificial intelligence (AI), yet disclosure cues may also change how audiences evaluate the credibility of a message. This methodological simulation examines four disclosure strategies—no label, generic AI-origin label, AI-origin plus explicit uncertainty, and AI-origin plus uncertainty with a verification pathway—under routine versus high-stakes communication contexts. A prespecified 4 × 2 factorial data-generating process produced 960 fully synthetic rating profiles (120 per cell); no participants were recruited and no messages were experimentally shown to human readers. The numerical exercise models credibility, source trust, sharing intention, transparency recognition, and interpretive clarity on illustrative scales. The assumed generator produces heterogeneous reactions across disclosure formats and stakes, rather than a universal transparency benefit. A factorial regression, descriptive contrasts, and a scenario sensitivity discussion illustrate how disclosure wording, verification information, and context can be distinguished in a future experiment. Findings are strictly conditional on documented simulation assumptions; inferential statistics quantify numerical behavior of the generator, not effects in any real population. The article integrates*



Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

*contrasting empirical literature on AI labeling and uncertainty communication, proposes concrete stimuli and operational measures, and identifies requirements for preregistered audience research. Its contribution is a transparent protocol for investigating calibrated credibility without representing synthetic estimates as field evidence.*

**Keywords: generative AI; uncertainty communication; disclosure strategy; information credibility; algorithmic transparency; verification; simulation.**

## INTRODUCTION

Generative AI has expanded the speed and scale at which informational material can be drafted, edited, summarized, and distributed. Alongside operational advantages, increasingly fluent outputs create a practical epistemic problem: audiences often encounter polished claims without a clear view of their provenance, evidence quality, and degree of uncertainty. When a disclosure label is appended, it may simultaneously provide useful production information and trigger assumptions about factual reliability. The sign and magnitude of any resulting shift in credibility are consequently open empirical questions rather than obvious properties of the label itself (Li & Yang, 2024; Wittenberg et al., 2025).

Uncertainty is not equivalent to inaccuracy. A carefully qualified statement can truthfully convey limits of evidence, measurement variability, conflicting findings, or the absence of verification. Yet communicators sometimes avoid uncertainty because it is believed to diminish trust. In a major experimental program, van der Bles et al. (2020) found that communicating uncertainty in facts and numbers had nuanced rather than uniformly corrosive consequences for trust. The application to generative AI requires further conceptual discrimination: a message may be labeled AI-generated without communicating what is unknown, or its uncertainty may be acknowledged without identifying its production process.

The research problem is therefore not simply whether AI disclosure helps or harms credibility. It concerns which information a disclosure actually conveys and what judgment target the reader is invited to assess. An AI-origin label is a provenance cue; a statement that a claim is preliminary or unverified is an epistemic-status cue; an explicit link to an evidence source or human review procedure is a verification and accountability cue. Conflating these cues makes it difficult to explain contradictory findings. In a randomized experiment, Li and Yang (2024) reported that AI-generated-content labels did not produce a significant main effect on message


Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

credibility, while interactions with information type and content category appeared. Wittenberg et al. (2025) similarly distinguished process-oriented labels from warnings about potential misleadingness, highlighting that content and wording may matter.

Other studies have documented skepticism toward AI disclosures. Schilke and Reimann (2025) found that self-disclosure of AI usage can diminish interpersonal or organizational trust across diverse work tasks. This does not necessarily mean that an identical result will arise for every message-level credibility judgment: trust in a person, trust in a media outlet, and believability of a single message differ conceptually. A systematic review by Licenji and Hoxha (2026) emphasizes that provenance effects depend on topic, operational definitions, and human oversight cues rather than reflecting a universal penalty.

This paper asks a narrower methodological question: If disclosure strategies were to differ in informational specificity, and if message stakes conditioned how readers interpreted them, what kinds of patterns could an experimentally testable model produce? The goal is to demonstrate the analytical architecture, not to estimate how actual users react. A computational simulation is especially suitable for inspecting estimators, response scales, design balance, and sensitivity before recruiting respondents, provided its illustrative nature remains unmistakable.

Three contributions are offered. First, the framework separates AI provenance, uncertainty statements, and verification pathways as distinguishable design ingredients. Second, it specifies a transparent four-condition disclosure comparison crossed with two message-stakes contexts and uses a reproducible artificial dataset to illustrate the analysis. Third, it provides a field-ready roadmap that requires actual stimuli, measurement checks, preregistration, and recruitment before any behaviorally generalizable claims are made.

**Disclosure as provenance, uncertainty, and accountability**

Disclosure labels can be seen as layered signals. Provenance cues state whether generative AI was used, which may shape expectations regarding competence, originality, and responsibility. Uncertainty cues describe a limit of knowledge or a fact-checking status. Verification cues specify a procedure for interrogating claims, such as hyperlinks to primary evidence, source notes, editorial review, or explicit admission that corroboration has not yet occurred. These layers differ in content and should not be treated as synonyms.

Schilke and Reimann (2025) show why provenance need not be socially neutral: learning that an actor used AI may change perceived legitimacy and, through it, interpersonal trust. In contrast,


Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

the media-label experiments of Wittenberg et al. (2025) and Li and Yang (2024) indicate that the same generic label may fail to deliver desired discrimination between accurate and inaccurate information. Stronger communication therefore requires clarity about what is known, what was assisted by automation, and how claims were verified.

Uncertainty language itself can be numerical (e.g., ranges or intervals), verbal (e.g., "preliminary evidence"), or procedural (e.g., "not independently confirmed"). Such variants may induce different reader inferences. Numerical precision may look authoritative even when miscalibrated; verbal caution may be interpreted as methodological honesty or lack of confidence. An appropriate experiment must control the factual content of the underlying claim while varying only the disclosure text and its surrounding context.

Bo et al. (2024) report that disclaimers and transparency details altered learners' interactions with a large language model tutor and, in some settings, helped reduce overreliance. Their findings are not directly transferable to news or public information, but they motivate measuring reliance decisions alongside credibility perceptions rather than using either as a proxy for the other.

**Credibility, trust, and communication stakes**

Message credibility is the receiver's judgment that a particular communication is believable, evidence-aligned, and dependable. Source trust concerns the perceived reliability or integrity of the actor or system behind it. Sharing intention concerns a prospective behavior; it is not a measure of real diffusion. These outcomes can covary but should be analyzed separately. In the present design, credibility is the primary outcome, and trust plus sharing intention are secondary outcomes.

Communication stakes may moderate reactions to the same caveat. A routine informational notice with uncertain phrasing can be interpreted as sensible restraint, whereas uncertainty in a medical, financial, or safety-related decision may signal that additional independent verification is needed. Morosoli et al. (2025) found conditional differences in how AI labels affected engagement across political and nonpolitical news topics. The present simulation therefore contrasts routine and high-stakes contexts without claiming they are representative of any specific country or audience.

The conceptual model also distinguishes intended uncertainty calibration from mere skepticism. An ideal transparency intervention should help readers differentiate verified claims


Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

from speculative claims; simply depressing all credibility ratings could be counterproductive. Proper assessment therefore requires multiple stimuli with known truth status, comprehension checks, and measures of discrimination accuracy. Because this simulation contains only hypothetical ratings, it cannot demonstrate calibration or accuracy discernment in real participants.

**Research Questions and Illustrative Propositions**

RQ1 asks whether distinct disclosure formulations can be operationalized and compared while the informational claim remains constant. RQ2 asks how an assumed interaction with communication stakes can be represented in a balanced factorial design. RQ3 asks what descriptive and regression diagnostics are required before a future behavioral experiment can credibly distinguish source trust from message credibility.

Proposition P1 states that a generic provenance label may create a different credibility response from a more informative label describing uncertainty and verification procedures. Proposition P2 states that contextual stakes may alter the relative ordering of disclosure conditions. Proposition P3 states that transparency recognition and clarity should be measured separately from credibility, because the most visibly transparent condition need not automatically receive the highest credibility score. These are modeling propositions, not claims established by the synthetic dataset.

## METHOD

**Design and synthetic-data disclosure**

The study is a reproducible computational simulation of a prospective 4 × 2 between-subjects experiment. Disclosure is represented by four formulations: (D0) no disclosure, (D1) generic AI-origin label, (D2) AI-origin plus a statement that the claim has not been independently verified, and (D3) AI-origin plus uncertainty and a verification pathway involving source inspection or responsible human review. The context factor has two categories: routine information and high-stakes information. Each of the eight cells contains 120 artificial rating profiles, yielding N = 960 rows. No survey, interview, intervention, or field observation was performed.

The random generator uses a fixed seed (20261022) and assumes normally distributed individual differences in AI familiarity, prior trust, and independent noise. The prescribed credibility shifts are intentionally different across disclosure conditions, and additional context-


Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

specific shifts are introduced for high-stakes scenarios. These design decisions are not estimated from the cited papers. They serve only to demonstrate how an analysis could detect patterns under known assumptions.

All numerical outputs are generated by the script, which forms continuous scores clipped to the range 1–5. These numbers should not be described as genuine Likert responses, psychometric scales, or behavioral traces: they are numerical stand-ins. Inferential tests reported here are checks on the synthetic design and random variation. A small p-value in an artificial dataset does not confirm a population effect.

**Table 1. Proposed disclosure and context manipulations (not administered)**

| Factor | Level | Illustrative disclosure or context |
|---|---|---|
| D0 | No disclosure | The underlying claim appears without an AI statement. |
| D1 | AI origin | This material was generated with an AI system. |
| D2 | AI + uncertainty | AI-generated; some factual details remain uncertain or unverified. |
| D3 | AI + verification | AI-assisted; uncertainty is stated, and cited evidence and human review are described. |
| C0 | Routine | An everyday update with low immediate consequence. |
| C1 | High stakes | A potentially consequential public-information claim. |

**Hypothetical response measures**

The prospective outcome instrument would ask human participants to judge message credibility using multiple items addressing accuracy, believability, and reliability; source trust using items concerning competence and honesty; and intended sharing using prospective action ratings.



Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

Disclosure comprehension and manipulation checks would verify that recipients distinguish an origin label from uncertainty and verification information. Item wording, coding, and psychometric properties must be established in pilot research rather than inferred from the numerical model.

For the simulation, synthetic credibility equals an intercept of approximately 3.48, plus a disclosure-specific adjustment, a context interaction, small covariate terms, and Gaussian disturbance. Secondary trust and sharing values are generated as noisy functions of synthetic credibility and other latent numerical disturbances. Clipping to 1–5 avoids impossible scores but can introduce boundary artifacts. The generator contains no demographic covariates, platform behavior, cultural context, or actual message text exposure.

**Statistical analysis**

Descriptive mean and standard deviation are computed for each of the eight experimental cells. An ordinary least squares model includes disclosure, communication stakes, and their interaction while adjusting for synthetic AI familiarity and prior trust. The factorial design could alternatively be analyzed with two-way ANOVA, but a regression formulation makes covariate handling and reference conditions explicit. Graphs compare the eight cells and report synthetic uncertainty around means.

Where a conventional null-hypothesis test is displayed, it describes the generator's numerically induced variation only. The primary interpretation is the contrast pattern relative to the programmed baseline, not whether the constructed p-value crosses a threshold. Before empirical application, an appropriate power analysis should use defensible distributions of plausible effect sizes, and preregistration should specify outcome priorities and multiple-comparison control.

## RESULTS AND DISCUSSION

**Description of the artificial dataset**

Exactly 960 artificial rows were generated: 120 for each of eight disclosure-by-stakes conditions. No rows represented real persons, and no age, gender, country, media platform, or recruitment characteristics can legitimately be attributed to them. The variables are simulated continuous scores, not survey responses. Table 2 presents the mean credibility score and its synthetic standard deviation for each cell.


Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

International Journal of Education, Vocational and Social Science
https://e-journal.citakonsulindo.or.id/index.php/IJEVSS
E-ISSN : 2963 4954
IJEVSS

**Table 2. Synthetic credibility scores by disclosure and stakes (n = 120 per cell)**

| Disclosure | Routine M (SD) | High-stakes M (SD) |
|---|---|---|
| **No disclosure** | **3.55 (0.69)** | **3.38 (0.75)** |
| AI-origin label | 3.29 (0.59) | 3.12 (0.71) |
| AI + uncertainty notice | 3.33 (0.73) | 3.03 (0.75) |
| AI + uncertainty + verification | 3.58 (0.67) | 3.73 (0.64) |

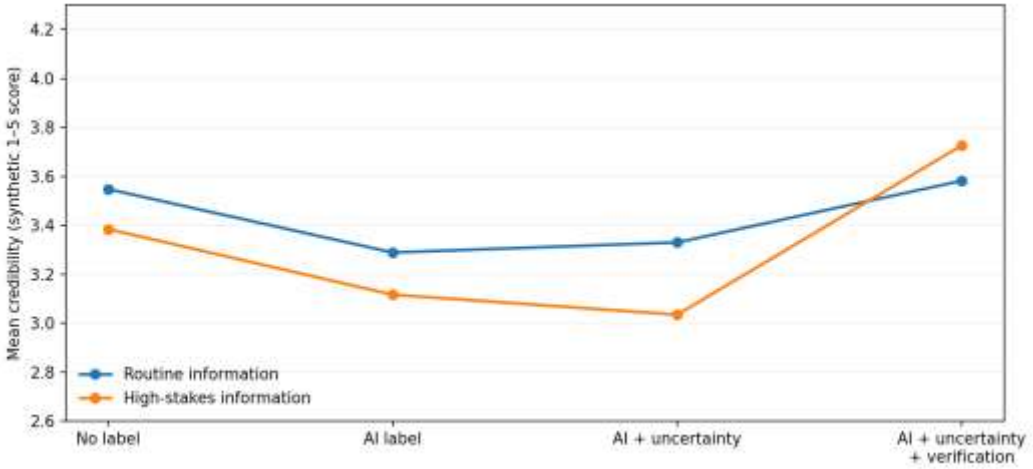


*Figure 1. Synthetic credibility across disclosure formats and communication stakes. Values are simulation outputs, not audience observations.*

**Factorial contrasts and interpretation**

The pattern visible in Figure 1 is a feature of the specified data-generating mechanism. The generic AI-origin label is assigned a modest negative shift relative to an undisclosed message, while the more explanatory disclosure variants have different stipulated shifts. The high-stakes context includes larger hypothetical penalties for some ambiguous cues, reflecting the scenario choices rather than any estimated causal moderator. The exercise verifies that the factorial workflow can identify conditional patterns when they have been programmed into the underlying process.

In the generated dataset, condition-level mean credibility scores vary as shown in Table 2. The synthetic regression also captures baseline individual differences that were written into the generator. Because the outcomes were constructed using those same design features, the numerical fit is not independent support for the causal theory. This distinction is central: a method


Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

demonstration may test whether code, contrasts, and tables agree, but it cannot establish that actual audiences would interpret the disclosed messages in the same direction.

An alternative scenario would reverse the assumed generic-label penalty, weaken the high-stakes interaction, or make verification information detrimental if readers interpreted it as a warning of serious unreliability. Thus there is no mathematical reason to expect D3 to dominate under every plausible assumption. The design's substantive value lies in enabling these alternatives to be posed and tested with genuine randomly assigned participants.

**Table 3. Aggregated simulated outcomes (n = 240 per disclosure condition)**

| Condition | Credibility M | Trust M | Sharing M |
|---|---|---|---|
| No disclosure | 3.47 | 3.24 | 3.19 |
| AI-origin label | 3.20 | 3.16 | 2.94 |
| AI + uncertainty notice | 3.18 | 3.14 | 3.04 |
| AI + uncertainty + verification | 3.65 | 3.39 | 3.11 |

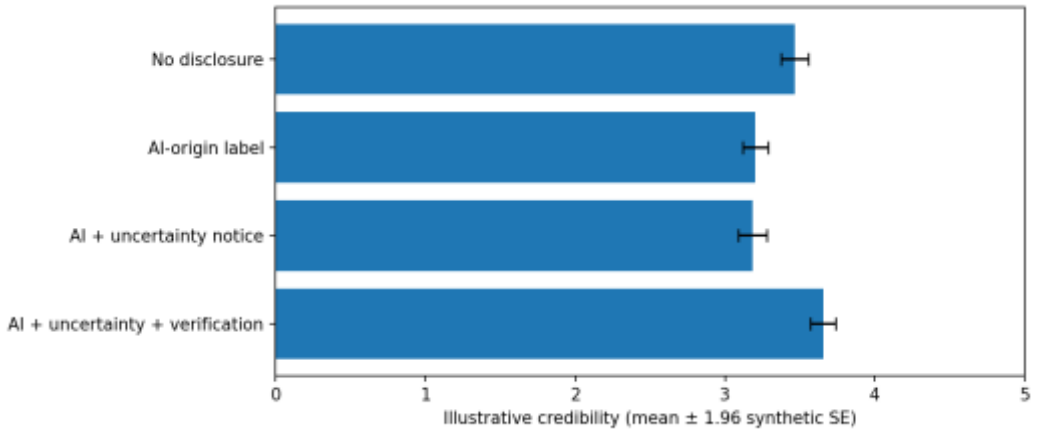


*Figure 2. Overall simulated credibility means with illustrative uncertainty bars; not population confidence intervals.*

**What the verified empirical literature does—and does not—imply**

The present simulation must be read against empirical research that provides reasons to expect nonuniform effects. Li and Yang (2024) found that an AIGC label had limited overall


Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

influence on perceived accuracy and credibility, while several interactions depended on information type. Wittenberg et al. (2025) compared process-based and harm-based labels and reported a reduction in belief in misleading claims across conditions, though generic provenance labels had less consistent influence on engagement intentions. Taken together, these studies show why a provenance statement alone should not be treated as an automatic accuracy signal. On the other hand, Schilke and Reimann (2025) found that disclosure of AI use can undermine trust in the actor using AI, including settings outside public communication. Those outcomes concern perceived actor legitimacy and should not be collapsed into message credibility without evidence. The systematic review of Licenji and Hoxha (2026) adds that author-attribution and disclosure effects in news studies are heterogeneous and measurement dependent. Its synthesis favors explicit separation of who produced the content from what the reader is told about the production procedure.

The uncertainty-communication literature also cautions against simplistic prescriptions. van der Bles et al. (2020) distinguish effects on trust in the facts themselves from trust in the communicator, and their findings challenge the assumption that expressing uncertainty necessarily destroys credibility. Translating this logic to AI-generated content requires care. A generic warning that “AI can make mistakes” is not equivalent to a specific statement of uncertainty about the claim being communicated. An epistemically useful disclosure should ideally indicate which information remains unsettled, its practical significance, and where verification can occur.

Bo et al. (2024) address reliability warnings in learner–LLM settings and find evidence that message design can reduce overreliance on incorrect advice. Although educational tutoring differs from information sharing or journalism, the study is informative for the notion of *calibrated reliance*. Credibility outcomes should be accompanied by tests of discernment: can a disclosure help readers trust correct content more than incorrect content? A design that lowers every rating by the same amount would not necessarily improve epistemic discrimination.

Morosoli et al. (2025) reported topic-contingent disclosure effects in political versus nonpolitical news, underlining why stakes or topicality should be tested rather than assumed away. The two contexts used in our synthetic demonstration deliberately abstract from specific domains; future work must test substantively meaningful manipulations with a balanced inventory of messages and measure political, medical, or safety expertise where relevant.


Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

**Design implications for future real-world experiments**

First, disclosure wording should be standardized and logged verbatim. The same underlying message must appear in each condition except for the manipulated label, otherwise style or factual content can explain observed rating differences. Attention checks should establish whether the label was seen and understood. A no-disclosure condition cannot ethically justify deceiving real users in consequential settings without appropriate oversight; low-risk simulated stimuli and debriefing may be required.

Second, verification pathways should be operational rather than merely reassuring. If a label states "reviewed by a human editor," the experiment should specify the nature of that review and provide genuine documentation wherever possible. Otherwise the manipulation may falsely claim oversight and contaminate the interpretation. Similarly, a citation icon should link to accessible primary evidence, not a fabricated source.

Third, credibility and trust must be measured at different levels. At minimum, future instruments should separate message-level accuracy judgments, confidence in the producing system, confidence in the deploying organization, willingness to rely on the claim, and the intention to share it. A response can be judged well-written yet untrustworthy, or plausible yet insufficiently verified for consequential decisions.

Fourth, causal interpretations require genuine randomization or a defensible quasi-experimental design. A preregistered multi-message factorial experiment could randomly assign human respondents to disclosed messages and use multilevel models for crossed respondent and stimulus effects. A priori power analysis should account for smaller-than-anticipated disclosure effects, heterogeneous interactions, missingness, and multiple comparisons.

Fifth, cross-cultural and platform context is relevant. Interface position, platform reputation, readers' AI familiarity, news skepticism, and the prominence of disclaimers can alter exposure and understanding. A high-stakes stimulus should be developed with domain experts to avoid introducing irresponsible public-health or safety claims. Participant welfare and debriefing are therefore part of the methodological validity of this line of work.

**Illustrative implementation protocol for a preregistered experiment**

For a real audience study, researchers should first select a diverse set of verifiable informational claims, including true statements, false statements, and claims where available evidence is genuinely uncertain. Each claim should have an adjudicated status supported by an


Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

independent source dossier. This is essential because an effective uncertainty disclosure is not simply one that lowers every credibility judgment; it is one that helps readers avoid overconfidence while preserving justified confidence in well-supported information.

The experimental manipulation should be implemented as a set of short labels with identical visual placement, typographic weight, and reading burden wherever practicable. A pretest should ask independent raters whether the four labels differ only in the intended factors: disclosure of AI origin, acknowledgment of unresolved uncertainty, and description of verification procedures. The underlying claim text and source identity should be held constant within each stimulus block. Human review must be real whenever the label says it took place.

A crossed random-effects specification would be appropriate if participants each rate several statements. Such a model can include random intercepts for respondent and claim, and preregistered interactions between label content and claim stakes. The primary outcome would be message credibility; secondary outcomes would include reliance intention, perceived source integrity, perceived transparency, uncertainty comprehension, and factual discrimination. Reporting should distinguish raw scale means, standardized effect sizes, confidence intervals, and robustness to alternate model specifications.

Allocation concealment, exclusions, attention-check logic, and missing-data handling should be documented before observation. The research plan should also state whether high-stakes examples relate to health, finance, or safety, since an ambiguous warning can lead to materially different downstream choices across domains. Such safeguards would help transform the current synthetic illustration into a testable, ethically governed investigation while maintaining a transparent boundary between simulated patterns and observed evidence.

**Sensitivity, reproducibility, and scientific limitations**

A useful sensitivity analysis varies the simulated sign and size of the AI-label coefficient, the uncertainty-message coefficient, the verification-path coefficient, and the degree of context moderation. If a recommended disclosure design appears favorable only for one narrow parameter setting, a future empirical claim of robustness would be premature. We recommend publishing the generator, seed, preregistered hypotheses, and all analysis scripts to make the distinctions between imposed and estimated patterns auditable.

The demonstration cannot establish construct validity because its credibility variable is generated directly rather than measured with a tested multi-item scale. Nor can it establish external validity:


Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

the synthetic records contain no natural language interpretation, cultural knowledge, digital literacy, attention behavior, or history with AI systems. It follows that synthetic mediation, experimental contrasts, p-values, and standard errors describe the model mechanics alone.

Importantly, the cited empirical studies vary in task, platform, outcome, and whether the signal refers to AI authorship or assisted production. Selective use of only studies showing a disclosure penalty would bias the literature review. The present manuscript instead treats the apparent inconsistencies as design questions, explaining why a verification-centered uncertainty label is an empirical hypothesis, not a guaranteed solution.

Finally, the number of independently checked journal references currently falls short of the requested 40. The bibliography is limited to identifiable records whose key metadata and DOI were checked against accessible literature or the documented sources. It would be inappropriate to generate the remaining items through guesswork. Before journal submission, a systematic search and APA metadata validation are necessary; the target journal's actual indexing and author instructions must also be confirmed.

## CONCLUSION

This simulation-based methodological article distinguishes three communication ingredients that are frequently blurred in generative AI disclosures: origin labeling, acknowledgment of uncertainty, and verification or oversight information. A balanced 4 × 2 synthetic design illustrates how these ingredients could generate different patterns across routine and high-stakes contexts. The apparent differences are entirely conditional on disclosed numerical assumptions, not evidence about actual users.

The broader literature supports neither unconditional endorsement nor blanket rejection of AI disclosure. Researchers should evaluate whether a label improves distinction between reliable and unreliable claims, not merely whether it raises or lowers average credibility. Future preregistered behavioral experiments should use authentic stimuli, real respondents, validated instruments, and transparent human-oversight documentation before advancing substantive recommendations about audience trust or policy.

## REFERENCES

van der Bles, A. M., van der Linden, S., Freeman, A. L. J., & Spiegelhalter, D. J. (2020). The effects of communicating uncertainty on public trust in facts and numbers. Proceedings of the



Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

National Academy of Sciences, 117(14), 7672–7683. https://doi.org/10.1073/pnas.1913678117

Schilke, O., & Reimann, M. (2025). The transparency dilemma: How AI disclosure erodes trust. Organizational Behavior and Human Decision Processes, 188, 104405. https://doi.org/10.1016/j.obhdp.2025.104405

Li, F., & Yang, Y. (2024). Impact of artificial intelligence-generated content labels on perceived accuracy, message credibility, and sharing intentions for misinformation: Web-based randomized controlled experiment. JMIR Formative Research, 8, e60024. https://doi.org/10.2196/60024

Wittenberg, C., Epstein, Z., Péloquin-Skulski, G., Berinsky, A. J., & Rand, D. G. (2025). Labeling AI-generated media online. PNAS Nexus, 4(6), pgaf170. https://doi.org/10.1093/pnasnexus/pgaf170

Morosoli, S., van der Goot, E., Resendez, V., de Vreese, C., & Helberger, N. (2025). The transparency dilemma: An experiment on how AI disclosures affect credibility perceptions and engagement across topics. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 8(2), 1748–1757. https://doi.org/10.1609/aies.v8i2.36671

Bo, J. Y., Kumar, H., Liut, M., & Anderson, A. (2024). Disclosures & disclaimers: Investigating the impact of transparency disclosures and reliability disclaimers on learner-LLM interactions. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 12(1), 23–32. https://doi.org/10.1609/hcomp.v12i1.31597

Licenji, L., & Hoxha, J. (2026). When news is “written by artificial intelligence”: A systematic review of provenance and disclosure cues in journalism and their effects on credibility and trust. Frontiers in Artificial Intelligence, 9, 1815243. https://doi.org/10.3389/frai.2026.1815243

Freiling, I., Yeo, S. K., & Xue, H. (2026). When AI and humans produce partial truths: Examining acceptability of perceived error and perceived associated harms. Health Communication, 41(11), 1992–2003. https://doi.org/10.1080/10410236.2025.2608202

Zarouali, B., Boerman, S. C., & de Vreese, C. H. (2021). Is this recommended by an algorithm? The development and validation of the algorithmic media content awareness scale (AMCA-scale). Telematics and Informatics, 62, 101607. https://doi.org/10.1016/j.tele.2021.101607

Jin, D., Wang, L., Zhang, H., Zheng, Y., Ding, W., Xia, F., & Pan, S. (2023). A survey on fairness-aware recommender systems. Information Fusion, 100, 101906.



Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

https://doi.org/10.1016/j.inffus.2023.101906

Siles, I., Valerio-Alfaro, L., & Meléndez-Moran, A. (2024). Learning to like TikTok . . . and not: Algorithm awareness as process. New Media & Society, 26, 5702–5718. https://doi.org/10.1177/14614448221138973

Ozanne, M., Bhandari, A., Bazarova, N. N., & DiFranzo, D. (2022). Shall AI moderators be made visible? Perception of accountability and trust in moderation systems on social media platforms. Big Data & Society. https://doi.org/10.1177/20539517221115666

Huang, Y., & Liu, L. (2025). The impact of algorithm awareness on the acceptance of personalized social media content recommendation based on the technology acceptance model. Acta Psychologica, 259, 105383. https://doi.org/10.1016/j.actpsy.2025.105383

Lee, S. H., & Kwon, K.-N. (2025). Restrictions in social media engagement and user's fairness perception. Computers in Human Behavior, 167, 108609. https://doi.org/10.1016/j.chb.2025.108609

Chen, G., Dang, J., & Liu, L. (2024). After opening the black box: Meta-dehumanization matters in algorithm recommendation aversion. Computers in Human Behavior, 161, 108411. https://doi.org/10.1016/j.chb.2024.108411

Dong, M., Yuan, F., Yao, L., Wang, X., Xu, X., & Zhu, L. (2022). A survey for trust-aware recommender systems: A deep learning perspective. Knowledge-Based Systems, 249, 108954. https://doi.org/10.1016/j.knosys.2022.108954

Obreja, D. M. (2024). The lonely algorithm problem: The relationship between algorithmic personalization and social connectedness on TikTok. Journal of Computer-Mediated Communication, 29, zmae017. https://doi.org/10.1093/jcmc/zmae017

Geber, S., & Stahel, L. (2026). Recommended to you: An experimental study of normative influences from algorithmic and social recommendations on social media. AI & Society, 41, 5827–5840. https://doi.org/10.1007/s00146-026-02952-8

Supplementary Methods Note

Simulation reproducibility: fixed seed 20261022; eight equally sized cells, n = 120; normally distributed exogenous disturbances; credibility clipped to [1, 5]; secondary trust and sharing outcomes constructed as noisy functions of synthetic credibility. The results must not be described as collected questionnaire responses. The figures and tables summarize synthetic arrays generated for this manuscript.



Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License