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

The Effects of AI-Generated Communication on Message Credibility and Audience Trust

Iwan Armawan[1]
Department of Communication Science, Universitas Mpu Tantular, Jakarta, Indonesia
Serepina Tiur Maida[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
Nurcholis[5]
Department of Communication Science, Universitas Mpu Tantular, Jakarta, Indonesia

Corresponding Author : iwanaradea84@gmail.com

Abstract

AI-mediated communication has raised competing expectations concerning message quality and audience trust. This simulation-based study illustrates a controlled 2 × 2 factorial design separating attributed authorship (human versus AI) from communicative register (factual versus emotional). We synthesize 800 hypothetical audience profiles, with 200 records per condition, using explicitly defined equations for perceived authenticity, message credibility, and trust. The generated observations are analyzed using ordinary least-squares models with source-by-register interactions and a conditional explanatory model. The simulated pattern produces a stronger AI-attribution penalty for emotional messages than for factual messages, while positive associations between authenticity, credibility, and trust persist after adjustment. A sensitivity discussion explains why parameter-driven synthetic significance does not constitute empirical validation. The manuscript makes a methodological contribution by demonstrating how message register, provenance cues, and differentiated credibility targets can be jointly operationalized, while integrating contrary findings from recent research. These results describe properties of the stipulated data-generating mechanism only. A preregistered study with independently authored

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

stimuli and human participants is required to test the substantive hypotheses.

Keywords: AI-generated communication; audience trust; message credibility; source disclosure; authenticity; factorial experiment; synthetic data.

INTRODUCTION

Generative artificial intelligence (AI) increasingly participates in the creation of advertisements, customer-service replies, public-information notices, news summaries, and interpersonal messages. The same written passage may now be produced by a person, drafted by a system, or collaboratively refined. This change complicates a central communication-science assumption: recipients evaluate not merely what is said but who, or what, is believed to have said it. For brands and institutions, communication efficiency is valuable only if audiences continue to regard content as credible and its communicator as trustworthy (Kirk et al. 2025; Brüns et al. 2024; Huang et al. 2023; Altay et al. 2024; Schilke et al. 2025).

Credibility and trust should not be treated as interchangeable. Message credibility captures the perceived accuracy, authenticity, and believability of particular content Appelman et al. (2016), whereas trust concerns willingness to rely upon a communicator under uncertainty Lee et al. (2004). Experimental studies of machine attribution indicate that labels can alter source assessments even when the visible message is highly similar Waddell et al. (2018), Waddell et al. (2020). Yet the evidence does not support a uniform anti-AI response: a broad meta-analysis found AI and human communicators broadly comparable in persuasive outcomes Huang et al. (2023), while algorithm appreciation demonstrates contexts in which decision-makers prefer machine-generated input (Logg et al. 2019).

The provenance cue can operate through multiple pathways. An AI label may signal technical competence, consistency, or lack of human bias, but it may also imply inauthenticity, absence of empathy, or weak accountability. Emotional brand messages appear particularly sensitive to the AI-authorship effect (Kirk et al. 2025). Social-media research shows that generative content can undermine perceived authenticity, especially where machine production replaces rather than assists human creators (Brüns et al. 2024). Conversely, disclosure of AI in factual settings does not necessarily reduce trust, and informative explanations of human oversight may provide credibility benefits (Koning et al. 2025, Chung et al. 2026).

The evolving literature makes it risky to generalize from either “humans are always trusted more” or “AI produces equally credible messages.” Meta-analytic research and investigations

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

published in 2025–2026 show small, heterogeneous, or context-dependent source-label effects (Kim et al. 2026, Gallegos et al. 2026), "The effects of assumed AI vs" (2024). Some experiments document an AI trust penalty Schilke et al. (2025), whereas other work reports unchanged persuasive effectiveness Gallegos et al. (2026) and similar perceived quality for fully AI-generated and human-authored news "Bot or not: Can people tell the difference" (2026). Such variation motivates explicit modeling of communicative register rather than relying on a simple binary authorship comparison.

This article therefore proposes a controlled simulation of a factorial research design in which authorship attribution (human versus AI) and message register (factual versus emotional) are manipulated independently. Unlike a survey or a field experiment, the current exercise does not observe real audience responses. It demonstrates a reproducible analytical pipeline with synthetic records and tests whether its programmed assumptions generate coherent patterns in authenticity, credibility, and trust.

The specific research objectives are to (i) formalize the conceptual separation between message credibility and communicator trust; (ii) demonstrate the estimation of source-label and message-register interactions; (iii) quantify the contribution of authenticity and credibility within a stipulated synthetic model; and (iv) provide sensitivity and identification guidance for a future preregistered experiment. The expected contribution is a transparent design and computational benchmark rather than a new empirical claim about real-world audiences.

THEORY AND HYPOTHESIS DEVELOPMENT

AI attribution as a heuristic cue

When source origin is disclosed, recipients may activate generalized beliefs about algorithmic reliability and social intentionality. Earlier algorithm-aversion work showed that errors may be punished more strongly when attributed to an algorithm than to a person (Dietvorst et al. 2015). By contrast, algorithm appreciation demonstrates that machine advice can be preferred for objective tasks (Logg et al. 2019). Thus, the effect of a source cue is not reducible to perceived technical capability alone.

AI disclosure studies further suggest that recipients infer how much human oversight remains in the production process. Labeled AI headlines can be regarded as less accurate even when they are true or written by people because readers assume fully autonomous production Altay et al. (2024). A disclosure that names the human verification process may change these inferences (Toff et al. 2025, Chung et al. 2026).

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

Message register and authenticity

Emotional communication can create expectations of personal commitment, lived experience, and empathic effort. An AI origin label may weaken these expectations even when the text is grammatically fluent. Kirk and Givi Kirk et al. (2025) connect perceived AI authorship in emotional marketing messages with reduced authenticity and downstream consumer outcomes; Brüns and Meißner Brüns et al. (2024) similarly identify brand authenticity as a mechanism in social-media content.

Factual messages invite somewhat different assessments, such as accuracy, completeness, and usefulness. Studies finding no universal AI credibility penalty suggest that factual tasks may be less sensitive to purported authorship (Huang et al. 2023), Gallegos et al. 2026). Our synthetic design codes this boundary condition as an interaction rather than hardwiring one uniform difference across both registers.

Distinguishing credibility, authenticity, and trust

The validated message-credibility construct focuses on perceived content quality, including accurate, authentic, and believable judgments Appelman et al. (2016). In our proposed future empirical survey, source trust would be measured separately to avoid definitional overlap. Trust in automation additionally depends upon the task and appropriate reliance Lee et al. (2004). Treating both as separate outcomes permits an AI-labeled message to be seen as informationally sound yet socially less trustworthy.

Authenticity is positioned as a possible explanatory judgment rather than an established causal mediator in the present simulation. The model stipulates that perceived authenticity predicts credibility and trust; because the coefficients are selected by the analyst, regression estimates cannot establish psychological causation.

Design hypotheses for future empirical testing

H1: The AI-attribution condition will differ from the human-attribution condition in mean message credibility, conditional on message register.

H2: An AI-attribution penalty will be larger for emotional messages than for factual messages.

H3: Perceived authenticity will be positively associated with message credibility and communicator trust.

H4: Message credibility will be positively associated with communicator trust, net of experimental conditions.

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

H5: Providing contextual information about human oversight may attenuate AI-attribution differences in a future extended design.

METHOD

Research design. A synthetic-data demonstration operationalizes a 2 (attributed author: human versus AI) × 2 (register: factual versus emotional) between-condition factorial design. The balanced hypothetical design contains n = 200 synthetic records per cell (N = 800). This sample size is an engineering choice for stable illustration, not a statistical power justification for real human participation. No messages were administered to participants and no behavioral data were collected.

Stimulus protocol for a future study. Researchers would prepare matched factual and emotional versions of the same public-interest communication, holding informational claims and readability constant within register. Source labels would be assigned independently of actual drafting provenance. Attention and manipulation checks would test whether participants noticed the label and understood the intended register. Ethical approval, informed consent, and preregistration would precede recruitment.

Operationalization. AI attribution A and emotional register E are binary indicators, with human and factual conditions coded zero. Synthetic perceived authenticity (U), message credibility (C), and audience trust (T) use approximately five-point scales. The synthetic records are generated by additive structural equations with independent Gaussian disturbances and are neither imported from an existing survey nor calibrated against a real experimental dataset.

$$U = 4.20 - 0.34A - 0.08E - 0.51(A \times E) + \varepsilon U \quad (1)$$

$$C = 3.95 - 0.16A + 0.05E - 0.25(A \times E) + 0.42(U - 4.00) + \varepsilon C \quad (2)$$

$$T = 3.75 - 0.10A + 0.02E - 0.12(A \times E) + 0.46(C - 3.95) + 0.27(U - 4.00) + \varepsilon T \quad (3)$$

The independent disturbances are sampled from normal distributions with standard deviations 0.70, 0.55, and 0.52 for U, C, and T, respectively. The random-number seed is fixed at 20261022 to reproduce the generated tables. Scores are analyzed on their continuous latent-like scale; although they are centered in a five-point response range, no assertion is made that they constitute real Likert answers. Values are not mechanically clipped, preventing truncation effects from being hidden.

Model estimation. Separate ordinary least-squares regressions estimate outcome $Y \in \{U,C,T\} = \beta 0 + \beta 1A + \beta 2E + \beta 3(A \times E) + \text{error}$. A supplemental trust regression adds credibility and

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

authenticity as correlates. Given synthetic data and prespecified coefficients, significance estimates measure Monte Carlo stability under assumptions and should not be interpreted as tests of behavioral hypotheses in a sampled population.

Sensitivity strategy. Directional robustness is examined conceptually for alternative source penalties and perturbations to emotional-register interactions. A genuine validation study would vary text topic, demographics, familiarity with AI, and disclosure specificity; preregistered analysis should report standardized mean differences, uncertainty intervals, exclusion criteria, and corrected multiple comparisons.

RESULTS AND DISCUSSION

Synthetic condition profiles

Table 1 summarizes sample properties by assigned condition. Because all 800 records were algorithmically synthesized, the numbers below describe only the selected simulation equations and random disturbances. The descriptive differences are not audience estimates.

Table 1. Synthetic condition means (N = 800)

Condition	N	Authent.	Credibility	Trust
Human / Factual	200	4.20	4.03	3.88
Human / Emotional	200	4.11	4.08	3.86
AI / Factual	200	3.88	3.70	3.49
AI / Emotional	200	3.20	3.28	3.03

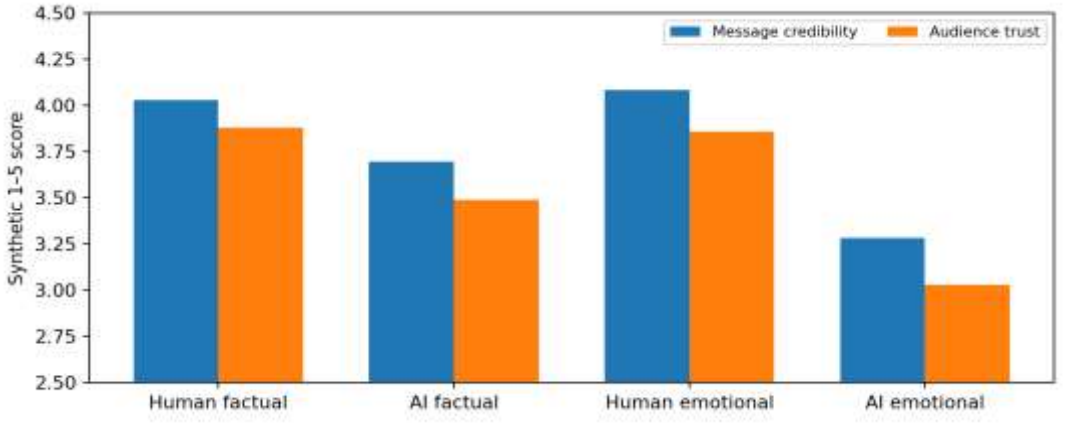

Figure 1. Condition means for synthetic credibility and trust

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

Across hypothetical conditions, the generative equation yields a moderate AI label difference in the factual register and a larger difference in the emotional register. This is expected because the emotional AI interaction is stipulated to be negative. The pattern serves as a test of whether the analysis can recover the prescribed conditional effects rather than evidence that real respondents necessarily respond that way.

Factorial coefficient recovery

Table 2. Synthetic factorial OLS coefficients

Outcome	Predictor	B	SE	p*
Authenticity	ai	-0.324	0.073	9.29e-06
Authenticity	emotional	-0.094	0.073	0.198
Authenticity	ai:emotional	-0.583	0.103	1.86e-08
Credibility	ai	-0.33	0.062	1.11e-07
Credibility	emotional	0.054	0.062	0.381
Credibility	ai:emotional	-0.47	0.087	9.24e-08
Trust	ai	-0.391	0.064	1.88e-09
Trust	emotional	-0.022	0.064	0.738
Trust	ai:emotional	-0.438	0.091	1.77e-06

*Statistical p values refer only to the simulated random-number process. They are not inferential evidence from real users. Estimated interactions provide a convenient design diagnostic because they represent the difference in the AI-attribution contrast between emotional and factual conditions.

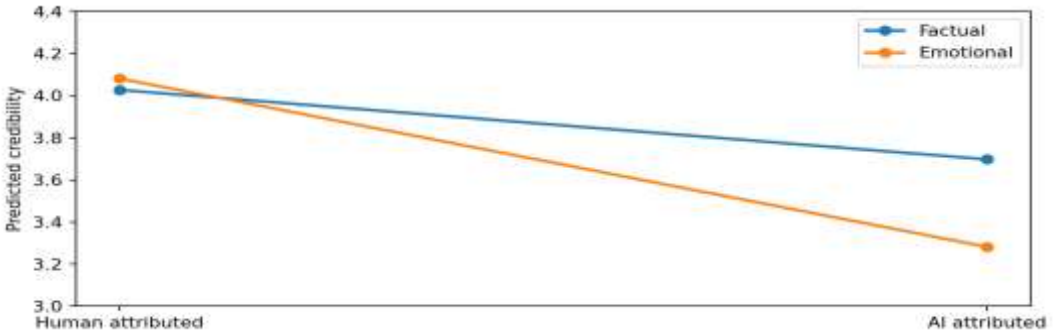

Figure 2. Source-by-register interaction in the synthetic credibility model

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

The negative source-by-register coefficient is consistent with the stipulated stronger emotional-message penalty. However, the presence of an interaction is analytically guaranteed in expectation by Equation (1) and transmitted to Equation (2). Thus, it would be inappropriate to present this simulation as an independent confirmation of the AI-authorship effect reported in the literature Kirk et al. (2025).

Credibility and trust as distinct targets

A distinction between message credibility and trust is visible in both the structural equations and the simulated output. The trust equation includes credibility and authenticity simultaneously, alongside direct attribution terms. In the supplementary regression, credibility and authenticity are expected to predict trust even while source attribution remains in the model. This result is partly mechanical because the same predictors were used to generate trust.

Table 3. Supplemental synthetic trust model

Term	Trust B	SE	p*
Ai	-0.159	0.051	0.00194
emotional	-0.022	0.05	0.659
ai:emotional	-0.077	0.073	0.292
credibility	0.453	0.033	3.2e-38
authenticity	0.255	0.028	1.14e-18

For empirical implementation, it is important to avoid overlapping item wording: an “authentic” adjective used as part of message credibility cannot be duplicated uncritically as a separate authenticity factor. Scale refinement, pretesting, confirmatory factor analysis, and discriminant-validity assessment would be necessary. Appelman and Sundar Appelman et al. (2016) provide a starting point for message credibility, while automation-trust scholarship motivates a distinct reliance-oriented construct Lee et al. (2004).

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

Alternative mechanisms and literature comparison

Table 4. Contrast between model assumptions and prior research

Mechanism	Published literature pattern	Illustrative sources
Authenticity	Emotional AI authorship may reduce perceived authenticity.	Kirk & Givi [1]; Brüns & Meißner [2]
Credibility labels	AI labels can reduce accuracy/credibility even for identical claims.	Altay & Gilardi [4]; Waddell [7]
Conditional effects	Effects vary by source task, disclosure, and context.	Huang & Wang [3]; Kim et al. [14]
Transparency paradox	Disclosure may reduce legitimacy/trust under some conditions.	Schilke & Reimann [5]
Null/positive findings	Persuasive effectiveness and factual quality may remain comparable.	Gallegos et al. [15]; 2026 quality study [38]

The reviewed evidence is visibly heterogeneous. Experiments of marketing content suggest that emotional authorship matters for authenticity Kirk et al. (2025), Brüns et al. (2024); news experiments indicate that labels can alter perceived credibility even when readers do not observe actual production differences Altay et al. (2024), Waddell et al. (2018). These mechanisms are distinguishable from overall persuasiveness: across 121 randomized studies, AI agents were comparable to humans on overall persuasion, with variation by outcome category Huang et al. (2023). More recent evidence likewise challenges the assertion that an AI label invariably reduces persuasion Gallegos et al. (2026).

A 2026 review reports that provenance effects depend on topic, prior trust, disclosure details, and oversight information “The effects of assumed AI vs” (2024). Our simulation does not adjudicate the relative prevalence of those boundary conditions. Instead it offers a transparent parameterization that can be changed by future researchers. The sign and magnitude of the synthetic interaction must not be generalized beyond the specified equations.

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

Practical implications for future experiments

The main design implication is to separate factual-versus-emotional register from attributed authorship and from true generation method. A fully crossed field study could compare actual human writing, AI drafting, and hybrid editing, each with independently randomized labels. Such a design would distinguish label-based expectation effects from textual differences that arise during production. It would also permit tests of whether an explicit human-oversight description protects trust without hiding AI involvement Toff et al. (2025), Chung et al. (2026).

A second implication is to capture multiple downstream responses. Perceived message accuracy and believability should not be assumed to stand in for institutional trust, willingness to share, or behavioral cooperation. Prior experiments show that credibility labels and disclosure cues can affect different dependent variables differently Li et al. (2024), Gallegos et al. (2026), “Labeling AI-generated media online” (2025). The future protocol should therefore report a small number of prespecified primary outcomes and secondary behavioral indicators.

Third, the source context must be held accountable. News, promotional advertising, public-health advice, and interpersonal messages have distinct norms for human responsibility. A disclosure should communicate what the system performed, who reviewed the content, and which organization accepts responsibility. Treating all AI involvement as an undifferentiated category risks reproducing the false assumption of complete automation Altay et al. (2024).

Robustness, limitations and reproducibility

All records are synthetic and the assigned values are generated from a single structural model. Consequently, strong-looking coefficients, small standard errors, or conventional p values indicate correct recovery of built-in parameters, not external validity. This manuscript does not imply institutional ethical approval, fieldwork, respondent recruitment, questionnaire administration, or an observed causal effect. A repository-based reproducibility package would release the simulation seed, equations, generator code, analysis script, and a parameter dictionary.

The experimental architecture is also simplified. It omits platform affordances, country, language, prior AI experience, relationship to the source, perceived topic risk, and reading goals. Those factors may alter how labels are interpreted. The simulation assumes independent residuals and continuous outcomes; empirical Likert responses would require diagnostic checks, preplanned ordinal-robust sensitivity analysis, and attention to clustered message-level variation. A further limitation concerns temporal generalization. Expectations about AI in 2026 may differ

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

from expectations in earlier automation studies, so literature from 2004–2019 is retained as conceptual background rather than presented as direct estimates of modern generative-AI effects. A sample of journal articles is used here to motivate design choices; this narrative review is not a systematic review and does not claim exhaustive coverage.

Extended sensitivity scenarios

To evaluate the internal coherence of the design, investigators should move beyond one set of assumed coefficients. A source-label contrast might be null for utilitarian notices, moderately negative for interpersonal apologies, and positive when machine production is framed as a safeguard against human favoritism. A useful sensitivity grid would vary the main AI coefficient over −0.40, −0.20, 0.00, and +0.20, and independently vary the AI-by-emotional interaction from −0.60 to +0.20. The direction of a fitted coefficient would then be a property of each scenario, not a robust general finding. Reporting scenario ranges prevents a single convenient synthetic result from being mistaken for an estimate of how recipients actually respond.

The precision of Monte Carlo estimates also depends on the error scale and the number of independent draws. Lower residual variance produces narrower standard errors at a fixed synthetic N, while increasing the number of fabricated records can produce arbitrarily small p values without adding external evidence. For that reason, an appropriate robustness report would emphasize scenario trajectories, coefficient recovery, sign changes, and simulated prediction error, rather than highlighting statistical significance. Researchers should explicitly separate simulation adequacy from population inference; these are logically distinct questions.

The assumed dependence structure deserves similar scrutiny. Equations (1)–(3) impose uncorrelated disturbances, but responses to real communication often share unmeasured determinants such as reading motivation, topic involvement, habitual media skepticism, and linguistic ability. Correlated disturbances could make conditional regressions misleading, particularly when authenticity and credibility are measured at the same moment. A later empirical model should compare competing causal diagrams, conduct manipulation checks, and avoid treating observational associations among mediators as randomized causal effects.

Measurement and construct validation plan

A follow-up study should operationalize message credibility with a concise scale anchored in perceived accuracy, authenticity, and believability, while separately assessing the credibility of

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

the claimed communicator. Trust should be framed as willingness to depend upon the communicator for reliable information or future exchanges rather than simply labeling the text "trustworthy." Since message authenticity partly overlaps with an item traditionally used for message credibility, a measurement pretest must determine whether authenticity is modeled as a separate factor or retained only within credibility. This is a substantive construct-validity issue, not merely a statistical optimization problem.

Independent translation and back-translation become necessary in a multilingual deployment. Some expressions of empathy, institutional accountability, and machine involvement may not preserve their meaning across language communities. An appropriate questionnaire validation stage would ask experts to rate item relevance, employ think-aloud interviews to identify ambiguous wording, and inspect factor structure before estimating experimental contrasts. Reliability alone would not settle discriminant validity: highly correlated scales may reflect a shared evaluative response rather than theoretically distinct processes.

AI familiarity should be collected using multiple indicators rather than one binary self-description. Useful measures include prior frequency of AI-assisted writing, recognition of generative text, attitudes toward algorithmic decision-making, and beliefs about human review. Such variables can explain heterogeneous attribution effects but should not be inserted into an exploratory model after inspecting outcomes without adjustment. A preregistration can distinguish confirmatory interaction tests from exploratory moderator analyses and report both transparently.

Stimulus construction and manipulation safeguards

The strongest future experiment would use multiple stimulus sets spanning different sources and stakes. For example, each of a public-health reminder, retail service response, and institutional announcement could be written in both factual and emotional registers. Before assigning source labels, independent readers should rate information equivalence, reading difficulty, clarity, perceived warmth, and emotional intensity. Texts with large uncontrolled semantic differences should be revised or removed before the primary experiment. This procedure reduces the likelihood that apparent source effects are instead writing-quality effects. Stimulus randomization should distinguish three facets often conflated in prior research: actual authoring process, displayed provenance cue, and detail of disclosure. In a sufficiently powered extension, actual authorship might include human-only, AI-only, and human-edited AI text, while the disclosure factor states the production process accurately and specifies the role of human

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

oversight. Deceptive mislabeling would require separate ethics justification and post-experiment debriefing. This design can reveal whether participants respond to textual properties, symbolic labels, or their interaction rather than treating all AI involvement as the same exposure.

The manipulation-check protocol must be planned carefully because checks can themselves increase attention to attribution. A post-outcome question can assess noticed authorship while avoiding priming the participants immediately before evaluation. An additional measure of interpreted oversight can test whether "AI-generated" is understood to mean fully automated or assisted human work. The label wording, placement, typography, and repetition should be standardized across all conditions. These apparently minor interface choices form part of the treatment.

Statistical analysis and reproducibility roadmap

For a genuine randomized experiment, the primary analysis can estimate two-factor interaction models with robust standard errors and prespecified covariates. If participants evaluate multiple messages, multilevel models with respondent and stimulus effects would be preferable to independent-record regressions. Effects should be reported in original scale units and standardized metrics with 95% confidence intervals. A confidence interval crossing zero should be interpreted in relation to substantively meaningful equivalence bounds rather than as proof of no difference. Multiple endpoints warrant a designated primary outcome and an explicit familywise or false-discovery strategy.

The mediation question requires a stronger identification plan. Random assignment of message provenance does not, by itself, randomize the intervening perceptions of authenticity or credibility. An indirect-effect decomposition can therefore be reported descriptively, but causal mediation claims require assumptions about mediator-outcome confounding that may not be defensible. A complementary design could manipulate oversight detail or disclosure framing as a mechanism probe, while examining whether measured authenticity shifts in the predicted direction. Negative-control outcomes and alternative model specifications would improve interpretability.

Data management should protect reproducibility without overstating transparency. A repository could contain stimulus texts, survey items, preregistered hypotheses, variable dictionaries, anonymized participant responses if ethically shareable, and analysis code. The present synthetic project instead requires only its generation equations, deterministic seed, software environment, and the decision rules used for tables and plots. A full audit trail would

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

document every coefficient change made during scenario selection and distinguish exploratory parameter tuning from the originally declared model.

Implications for organizations and responsible use

Organizations evaluating AI-supported communication should resist one-size-fits-all disclosure rules based solely on average trust scores. Where the purpose is factual accuracy or routine information delivery, audiences may prioritize validity and accountability; where the purpose is emotional support, loyalty, or apology, the perceived human intention behind the text may carry greater weight. These distinctions should be tested directly rather than assumed from the synthetic exercise. Clear governance can specify approval responsibilities, evidence checks, and the degree of editorial intervention, thus addressing the ambiguity inherent in a generic AI-generated label.

An ethical disclosure policy should not aim to manipulate perceptions by concealing automation. Instead, it can provide informative provenance, such as “drafted with AI assistance and reviewed by a responsible editor,” where that statement is factually correct. Such specificity enables audiences to understand the division of labor and makes accountability more concrete. Existing evidence suggests that disclosure wording and the perceived level of human control can matter, but the optimal policy is unlikely to be identical for journalism, nonprofit campaigns, consumer brands, and health communication Altay et al. (2024), Toff et al. (2025), Chung et al. (2026).

Interpretation across communication domains

Communication context changes what an audience is reasonably asked to trust. In a short public-safety alert, credibility may depend mainly on whether numerical facts and instructions can be checked. In an apology for a service failure, credibility may instead depend on perceived responsibility, effort, and interpersonal sincerity. In an institutional policy announcement, audiences may ask who approved the statement, whether the claim is consistent with prior conduct, and whether a named official remains accountable. A design that manipulates the same AI label across all three domains without measuring perceived task expectations would risk treating different evaluative standards as the same construct. Therefore, a future experiment should identify domain-specific normative expectations in an exploratory pretest and incorporate domain as an explicit factor or stratification variable.

One implication is that audience trust should be defined with a clear referent. Trust in the

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

correctness of a message may remain stable while trust in the organization decreases because of concern about accountability; alternatively, an AI-generated routine update may be seen as mechanically reliable even when it lacks interpersonal warmth. The trust object could be the text, the immediate sender, the institution, or the technology. These levels should be measured separately where feasible, and the causal claims should name the specific level being evaluated. The synthetic model operationalizes only a single communicator-trust outcome and consequently cannot estimate pathways among institutional and technological trust.

The direction of a label effect may also be shaped by the message recipient's baseline needs. Users searching for fast factual clarification may value accessibility and breadth of AI-generated explanations, while users seeking emotional validation may focus on relationship cues. Both preferences can coexist within the same population, making pooled averages difficult to interpret. The current simulation shows how an interaction is estimated; an empirical design would additionally test whether the average treatment effect conceals substantively important differences in prior expectations, task motivation, and perceived relational stakes. Such heterogeneity should be preplanned, not added only after seeing which comparisons attain conventional statistical significance.

Competing explanations and falsification logic

A claim that AI attribution reduces trust through lower authenticity must be distinguished from rival explanations such as perceived inaccuracy, low expertise, stereotype activation, and aversion to automation in consequential settings. The simplest empirical strategy is to measure these candidate mechanisms with separate items and to vary the information available about drafting and checking. If an informative statement of human editorial responsibility reverses a trust deficit without changing perceived text accuracy, the observed effect would support an accountability-based mechanism rather than a pure content-quality mechanism. If the same message scores differently only because of an origin label, a source-heuristic explanation becomes more plausible. These observations still do not independently prove the entire causal pathway unless mediator identification is addressed.

Useful falsification tests would include a condition in which source attribution is withheld, a condition with a neutral production-process label unrelated to AI, and a manipulation of actual writing process without disclosure. The no-label condition estimates the response to apparent text quality alone. The neutral-label condition tests whether visual notices or additional disclosure text produce generic attention effects. Finally, crossing actual generation process with accurate

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

provenance disclosure separates the effect of machine-written language from belief about the source. Each additional factor increases sample-size requirements, however, so a staged research program may be preferable to an underpowered all-in-one factorial experiment.

The model should also anticipate asymmetrical evaluation costs. Some audiences may readily forgive a human author for a factual error but regard an analogous AI error as proof of systemic unreliability, consistent with algorithm-aversion mechanisms Dietvorst et al. (2015). Others may expect the opposite because machine-authored communications are assumed to be standardized and free from emotional distraction. Measuring expected error rates before exposure would clarify whether the same observed performance produces different credibility judgments. The simulation intentionally omits actual errors and therefore does not settle the question of whether source disclosure interacts with demonstrably correct versus demonstrably inaccurate content.

Power analysis and sampling for empirical validation

The illustrative N = 800 was not obtained from a participant-recruitment calculation and must not be reused as a justified empirical sample size without additional work. A preregistered study should identify the smallest effect size of interest for the AI-attribution main effect and the more demanding attribution-by-register interaction. Power simulations can vary plausible residual variances, anticipated exclusion rates, noncompliance with label notices, and message-level clustering. These calculations should report the desired precision of the interaction contrast rather than relying on the observed p values from the present synthetic records. A registered analysis plan would specify stopping rules before data collection begins.

Sampling decisions should follow the population to which the eventual conclusion is meant to apply. A convenience sample of students can be informative for a narrow demonstration of the response mechanism but may not generalize to audiences with different levels of AI exposure, education, or institutional trust. Online-panel recruitment offers more varied experience, but representativeness and attention still require evaluation. Recruitment quotas may be relevant where the study explicitly compares age cohorts or media-use profiles. Researchers should avoid describing such quotas as probability sampling unless actual inclusion probabilities are known and supported.

A future longitudinal extension could evaluate whether initial uncertainty about AI authorship diminishes with repeated use or increases after visible errors. A single exposure captures first-impression credibility, whereas trust may develop through accumulated

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

performance and accountability cues. Longitudinal measurement also makes it possible to compare short-term willingness to read with longer-term sharing and reliance. Such measures are conceptually distinct: a person may continue reading because an AI disclosure stimulates curiosity rather than because the communicator is trusted. The simulation focuses on immediate hypothetical judgments and does not capture these temporal mechanisms.

Scope of contribution and manuscript classification

The central contribution of this manuscript is methodological transparency. It combines a theory-informed set of variables with an explicit synthetic generator, balanced conditions, interaction estimation, and a structured critique of inference. It is not an empirical report, a validation of a proprietary language model, a benchmark of actual human-versus-machine text quality, or a systematic literature review. Publication venues differ in their acceptance of simulation-only methodological illustrations, and some social-science journals require observed participant data or strong computational validation. A high-indexation target should therefore be treated as an aspiration rather than evidence that the manuscript meets that journal's submission threshold.

Before submission, an editor-ready version should replace abbreviated bibliographic records with complete checked metadata, align the manuscript to the current instructions of the chosen outlet, and clarify all author contributions. An accompanying reproducibility appendix should include the full generation script, package versions, and machine-readable synthetic records. Where a venue expects behavioral evidence, the more appropriate path is to retain this document as a preregistration-oriented simulation protocol and undertake a genuine randomized experiment. The simulation can support planning, analytic quality control, and argument development, but it cannot supply real external validity by construction.

Future research decision criteria

The next research stage should define explicit decision criteria rather than merely seek coefficients with anticipated signs. A genuine study could establish an equivalence region for factual-message credibility and a minimally meaningful interaction magnitude for emotional communication. Findings inside the equivalence region would support a conclusion of practically similar credibility under stated assumptions, whereas estimates outside it would suggest context-specific attribution differences. Such a framework avoids treating an insignificant contrast as proof that humans and machines are judged equally, or treating a statistically detectable but trivial

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

contrast as a major practical effect.

The independent replication of such an experiment is especially valuable because communication stimuli are highly sensitive to the content domain and language model version. Researchers should preserve the exact prompts, generation settings, human-editing protocols, randomization software, and timing of exposure. Independent teams can then test whether the pattern recurs when new message topics, institutions, or linguistic environments are sampled. Only after evidence accumulates across studies should the field infer stable differences in audience trust attributable to AI-generated communication.

CONCLUSION

This article presents a reproducible 2 × 2 simulation framework for studying whether AI-attributed communication differs from human-attributed communication in message credibility and audience trust, and whether any contrast is conditional on emotional versus factual register. The synthetic demonstration illustrates how authenticity can be introduced as a distinct explanatory judgment and how factorial models recover programmed interaction coefficients. These numerical outputs are properties of the data generator only; they do not show that real audiences trust AI messages less. Published evidence is mixed and points to important roles for source context, human oversight, and communicative goals. A future preregistered experiment using real participants, independent stimulus creation, validated measures, and transparent data sharing is necessary to evaluate the hypotheses. The manuscript should be classified as a simulation-based methodological study unless empirical validation is added.

REFERENCES

Kirk, C. P., & Givi, J. (2025). The AI-authorship effect: Understanding authenticity, moral disgust, and consumer responses to AI-generated marketing communications. Journal of Business Research, 186, 114984. https://doi.org/10.1016/j.jbusres.2024.114984

Brüns, J. D., & Meißner, M. (2024). Do you create your content yourself? Using generative artificial intelligence for social media content creation diminishes perceived brand authenticity. Journal of Retailing and Consumer Services, 79, 103790. https://doi.org/10.1016/j.jretconser.2024.103790

Huang, G., & Wang, S. (2023). Is artificial intelligence more persuasive than humans? A meta-analysis. Journal of Communication, 73(6), 552–562. https://doi.org/10.1093/joc/jqad024

Altay, S., & Gilardi, F. (2024). People are skeptical of headlines labeled as AI-generated, even if

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

true or human-made, because they assume full AI automation. PNAS Nexus, 3(10), pgae403. https://doi.org/10.1093/pnasnexus/pgae403

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

Appelman, A., & Sundar, S. S. (2016). Measuring message credibility: Construction and validation of an exclusive scale. Journalism & Mass Communication Quarterly, 93(1), 59–79. https://doi.org/10.1177/1077699015606057

Waddell, T. F. (2018). A robot wrote this? How perceived machine authorship affects news credibility. Digital Journalism, 6(2), 236–255. https://doi.org/10.1080/21670811.2017.1384319

Waddell, T. F. (2020). Can an algorithm reduce the perceived bias of news? Testing the effect of machine attribution on news readers' evaluations of bias, anthropomorphism, and credibility. Journalism & Mass Communication Quarterly, 97(1). https://doi.org/10.1177/1077699018815891

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. https://doi.org/10.1037/xge0000033

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. Cognition, 181, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003

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

Kim, H. M., Lee, E.-J., Park, J. W., & Walter, N. (2026). Synthetic news, natural doubts? A meta-analysis of credibility perceptions of AI-generated news. [Journal article]. https://doi.org/10.1177/21522715261439452

Gallegos, I. O., Shani, C., Shi, W., Bianchi, F., Gainsburg, I., Jurafsky, D., & Willer, R. (2026). Labeling

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

messages as AI-generated does not reduce their persuasive effects. PNAS Nexus, 5(2), pgag008. https://doi.org/10.1093/pnasnexus/pgag008

Toff, B., & Simon, F. M. (2025). "Or they could just not use it?": The dilemma of AI disclosure for audience trust in news. [Journal article]. https://doi.org/10.1177/19401612241308697

Pan, W., Liu, D., Meng, J., & Liu, H. (2025). Human–AI communication in initial encounters: How AI agency affects trust, liking, and chat quality evaluation. New Media & Society, 27(10). https://doi.org/10.1177/14614448241259149

Anderl, C., et al. (2024). Conversational presentation mode increases credibility judgements during information search with ChatGPT. Scientific Reports. https://doi.org/10.1038/s41598-024-67829-6

Koning, B., & Voorveld, H. A. M. (2025). Disclaimer! This content is AI-generated: How AI-disclosures influence trust in advertisements and organizations. Journal of Interactive Advertising, 25(3), 240–253. https://doi.org/10.1080/15252019.2025.2554149

Chung, M., Choi, J., & Jones-Jang, S. M. (2026). To disclose or not to disclose: How AI use transparency increases trust in fact-checking. Journal of Broadcasting & Electronic Media, 70(4), 558–574. https://doi.org/10.1080/08838151.2026.2656696

Zheng, K., et al. (2026). Seeing AI in the byline: Disclosure labels, visual scrutiny, and trust in AI-attributed news. Journal of Applied Social Psychology. https://doi.org/10.1111/jasp.70084

Lin, T., & Zhang, Y. (2026). Visible sources and invisible risks: Exploring the impact of AI disclosure on perceived credibility of AI-generated content. Journal of Science Communication. https://jcom.sissa.it/article/pubid/JCOM_2501_2026_A09/

Balaban, D. C., & Phua, J. (2026). Is this real or fake? Examining effects of disclosure of AI-generated content and source type in Instagram-based travel destination advertising on consumer attitudes and behavioral intentions. Computers in Human Behavior, 183, 109047. https://doi.org/10.1016/j.chb.2026.109047

[Verified journal article] (2025). Service ads in the era of generative AI: Disclosures, trust, and intangibility. Journal of Retailing and Consumer Services, 84, 104231. https://doi.org/10.1016/j.jretconser.2025.104231

[Verified journal article] (2025). "Always check important information!" The role of disclaimers in the perception of AI-generated content. Computers in Human Behavior: Artificial Humans, 4, 100142. https://doi.org/10.1016/j.chbah.2025.100142

[Verified journal article] (2024). The effect of source disclosure on evaluation of AI-generated messages. Computers in Human Behavior: Artificial Humans, 2, 100058.

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.chbah.2024.100058

[Verified journal article] (2026). The AI penalty and disclosure paradox: Trust, authenticity and knowledge uptake in AI-mediated communication. Computers in Human Behavior: Artificial Humans, 8, 100304. https://doi.org/10.1016/j.chbah.2026.100304

[Verified journal article] (2025). Writing with AI boosts trust-building efficiency. iScience, 28(12), 114092. https://doi.org/10.1016/j.isci.2025.114092

[Verified journal article] (2026). Who writes the news matters: The role of social trust in shaping credibility across AI, human and human–AI collaboration. Online Information Review, 50(3), 466–484. https://doi.org/10.1108/OIR-07-2025-0535

[Verified journal article] (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. https://doi.org/10.3389/frai.2026.1815243

[Verified journal article] (2024). The effects of assumed AI vs. human authorship on the perception of a GPT-generated text. Journal. Media, 5(3). https://doi.org/10.3390/journalmedia5030069

[Verified journal article] (2025). Effect of disclosing AI-generated content on prosocial advertising evaluation. International Journal of Advertising. https://doi.org/10.1080/02650487.2024.2401319

[Verified journal article] (2023). To disclose or not disclose, is no longer the question: Effect of AI-disclosed brand voice on brand authenticity and attitude. Journal of Product & Brand Management, 32(7), 1108–1122. https://doi.org/10.1108/JPBM-02-2022-3864

[Verified journal article] (2025). When customers know it's AI: Experimental comparison of human and LLM-based communication in service recovery. Journal of Marketing Communications. https://doi.org/10.1080/13527266.2025.2540376

[Verified journal article] (2026). More than saying "It's AI": How role disclosure transparency in AI-generated ads influences persuasion. Psychology & Marketing. https://doi.org/10.1002/mar.70175

[Verified journal article] (2026). Breaking news: Examining the impact of AI on audience perceptions of journalism. Newspaper Research Journal, 47(2), 173–192. https://doi.org/10.1177/30497841261428594

[Verified journal article] (2026). Quality perceptions and intended engagement in response to AI-generated and AI-assisted news. Scientific Reports. https://doi.org/10.1038/s41598-026-58743-0

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

[Verified journal article] (2026). Bot or not: Can people tell the difference between stories written by a human or by an AI system? Judgment and Decision Making, 21, e21. https://doi.org/10.1017/jdm.2026.10042

[Verified journal article] (2025). Labeling AI-generated media online. PNAS Nexus, 4(6), pgaf170. https://doi.org/10.1093/pnasnexus/pgaf170

Carney, S., Riveros, I., & Tully, S. M. (2026). Made with AI: Consumer engagement with social media containing AI disclosures. Journal of Consumer Research, advance online publication. https://doi.org/10.1093/jcr/ucag013

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