p-ISSN:2708-2091
e-ISSN:2708-3586

www.gsrjournal.com
GSR
Global Sociological Review

GSR

GLOBAL SOCIOLOGICAL REVIEW

HEC-RECOGNIZED CATEGORY-Y

VOL. X ISSUE III, SUMMER (SEPTEMBER-2025)

Double-blind Peer-review Research Journal
www.gsrjournal.com

DOI (Journal): 10.31703/gsr
DOI (Volume): 10.31703/gsr.2025(X)
DOI (Issue): 10.31703/gsr.2025(X-III)

HumanityPublications
sharing research

Humanity Publications (HumaPub)
www.humapub.com
Doi: https://dx.doi.org/10.31703

Article title

The Impact of Artificial Intelligence on News Credibility and Media Ethics

Abstract

The use of artificial intelligence (AI) in newsrooms creates the issue of how it will affect perceived news credibility and ethical standards. The online experiment (n=1,200) was conducted in a 2x2 format and investigated the impact of message credibility, accuracy, fairness, and sharing willingness between byline (Human vs. AI-assisted) and disclosure (None vs. Explicit). The findings indicated a low but significant penalty against AI-assisted bylines, which was alleviated by direct disclosure, suggesting editorial control (AI-assisted; editor verified), particularly in people more media literate. The qualitative results indicated that the human-in-the-loop verification, the division of tasks, and the provenance practices could tackle these issues. The paper recommends that responsible AI usage involves a set of dedicated disclosure, human validation, and transparency to ensure credibility and ethical standards.

Keywords: Artificial Intelligence; News Credibility; Media Ethics; Disclosure; Human-In-The-Loop; Provenance; Algorithmic Accountability

Authors:

Robina Saeed: (Corresponding Author)
Associate Professor, School of Media and Communication Studies, Minhaj University, Lahore, Punjab, Pakistan.
(Email: robina.mcomm@mul.edu.pk)

Maryam Hashmi: PhD scholar, Department of Media and Communication Studies, International Islamic University, Islamabad, Pakistan.

Saadia Qamar: M.Phil, School of Media and Communication Studies, Minhaj University, Lahore, Punjab, Pakistan.

Pages: 90-108
DOI: 10.31703/gsr.2025(X-III).10
DOI link: https://dx.doi.org/10.31703/gsr.2025(X-III).10
Article link: http://www.gsrjournal.com/article/the-impact-of-artificial-intelligence-on-news-credibility-and-media-ethics
Full-text Link: https://gsrjournal.com/article/the-impact-of-artificial-intelligence-on-news-credibility-and-media-ethics
Pdf link: https://www.gsrjournal.com/jadmin/Auther/31rvIolA2.pdf

Global Sociological Review
p-ISSN: 2708-2091 **e-ISSN:** 2708-3586
DOI(journal): 10.31703/gsr
Volume: X (2025)
DOI (volume): 10.31703/gsr.2025(X)
Issue: III Summer (September-2025)
DOI(Issue): 10.31703/gsr.2024(X-III)

Home Page
www.gsrjournal.com

Volume: (2025)
https://www.gsrjournal.com/Current-issues

Issue: III-Summer (June -2025)
https://www.gsrjournal.com/issue/10/3/2025

Scope
https://www.gsrjournal.com/about-us/scope

Submission
https://humaglobe.com/index.php/gsr/submissions

Humanity Publications (HumaPub)
www.humapub.com
Doi: https://dx.doi.org/10.31703

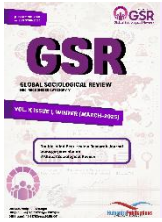
Citing this Article

10	The Impact of Artificial Intelligence on News Credibility and Media Ethics		
Authors	Robina Saeed Maryam Hashmi Saadia Qamar	DOI	10.31703/gsr.2025(X-III).10
		Pages	90-108
		Year	2025
		Volume	X
		Issue	III

Referencing & Citing Styles

APA	Saeed, R., Hashmi, M., & Qamar, S. (2025). The Impact of Artificial Intelligence on News Credibility and Media Ethics. *Global Sociological Review*, *X*(III), 90-108. https://doi.org/10.31703/gsr.2025(X-III).10
CHICAGO	Saeed, Robina, Maryam Hashmi, and Saadia Qamar. 2025. "The Impact of Artificial Intelligence on News Credibility and Media Ethics." *Global Sociological Review* X (III): 90-108. doi: 10.31703/gsr.2025(X-III).10.
HARVARD	SAEED, R., HASHMI, M. & QAMAR, S. 2025. The Impact of Artificial Intelligence on News Credibility and Media Ethics. *Global Sociological Review,* X**,** 90-108.
MHRA	Saeed, Robina, Maryam Hashmi, and Saadia Qamar. 2025. 'The Impact of Artificial Intelligence on News Credibility and Media Ethics', *Global Sociological Review*, X: 90-108.
MLA	Saeed, Robina, Maryam Hashmi, and Saadia Qamar. "The Impact of Artificial Intelligence on News Credibility and Media Ethics." *Global Sociological Review* X.III (2025): 90-108. Print.
OXFORD	Saeed, Robina, Hashmi, Maryam, and Qamar, Saadia (2025), 'The Impact of Artificial Intelligence on News Credibility and Media Ethics', *Global Sociological Review,* X (III), 90-108.
TURABIAN	Saeed, Robina, Maryam Hashmi, and Saadia Qamar. "The Impact of Artificial Intelligence on News Credibility and Media Ethics." *Global Sociological Review* X, no. III (2025): 90-108. https://dx.doi.org/10.31703/gsr.2025(X-III).10.

e-ISSN: 2708-2091 Volume: X (2025) Issue: III-Summer (September -2025) p-ISSN: 2708-3586



Global Sociological Review
www.gsrjournal.com
DOI: http://dx.doi.org/10.31703/gsr

Pages: 90-108 URL: https://doi.org/10.31703/gsr.2025(X-III).10 Doi: 10.31703/gsr.2025(X-III).10

Title

The Impact of Artificial Intelligence on News Credibility and Media Ethics

Authors:

Robina Saeed:
(Corresponding Author)
Associate Professor, School of Media and Communication Studies, Minhaj University, Lahore, Punjab, Pakistan.
(Email: robina.mcomm@mul.edu.pk)
Maryam Hashmi:
PhD scholar, Department of Media and Communication Studies, International Islamic University, Islamabad, Pakistan.
Saadia Qamar:
M.Phil, School of Media and Communication Studies, Minhaj University, Lahore, Punjab, Pakistan.

Contents

Abstract

The use of artificial intelligence (AI) in newsrooms creates the issue of how it will affect perceived news credibility and ethical standards. The online experiment (n=1,200) was conducted in a 2x2 format and investigated the impact of message credibility, accuracy, fairness, and sharing willingness between byline (Human vs. AI-assisted) and disclosure (None vs. Explicit). The findings indicated a low but significant penalty against AI-assisted bylines, which was alleviated by direct disclosure, suggesting editorial control (AI-assisted; editor verified), particularly in people more media literate. The qualitative results indicated that the human-in-the-loop verification, the division of tasks, and the provenance practices could tackle these issues. The paper recommends that responsible AI usage involves a set of dedicated disclosure, human validation, and transparency to ensure credibility and ethical standards.

Keywords: Artificial Intelligence; News Credibility; Media Ethics; Disclosure; Human-In-The-Loop; Provenance; Algorithmic Accountability

Introduction

Artificial intelligence (AI) is now infiltrating every step of the news pipeline, including its collection, writing, editing, optimization of the headline, recommendation, and editing, potentially bringing more efficiencies than at any time before it, but also

exacerbating historical issues of misinformation and a decrease in trust in journalism (Cheong, 2024; Sonni et al., 2024). At the newsroom, AI is increasingly implemented in the background, often on small-scale tasks, like document summarization or SEO headline creation, as well as becoming increasingly apparent as a means of content creation and delivery via algorithmic bylines and recommender systems (Cools & Koliska, 2024; Mitova et al., 2024). However, the responses of the public are two-sided: viewers fear that AI kills authenticity and accountability in the areas where the value of verification and ethical decisions is defined the most, politics, health, crises, etc. (Lundberg, 2024). The ensuing credibility dilemma poses a fundamental ethical concern: can the achievability of AI efficiency be balanced with the ethics of journalism to tell the truth, disclose the truth, and be responsible?

The recent empirical studies in the last several years have started mapping the influence of AI cues on the audience's judgment of news credibility. The findings of experiments have shown that merely stating that AI is involved can lower perceived trustworthiness, independent of whether the article discussed is the same, which is evidence of a transparency penalty in certain situations (Toff & Simon, 2024; Schilke & Reimann, 2025). The associated research in the area of bylines indicates that perceived machine authorship is a negative source of credibility of sources and messages, and that the effect is mediated by the judgment of humanness (Jia et al., 2024). These results make it difficult to accept the standard wisdom that more disclosure is always better to promote trust; rather, it depends on the manner in which disclosures are packaged and the supplementary quality-assurance signals (e.g., source used) attached to them (Toff & Simon, 2024).

Within news organizations, AI changes ethical work as a governance issue, rather than an issue of perception by the audience. Examples of large outlets indicate that both internal and external transparency might differ: engineering teams can document or describe some of the automated systems, but newsroom communication with journalists and the general public is incomplete, haphazard, or excessively technical (Cools & Koliska, 2024). Meanwhile, media ethics and AI governance scholarship highlight that transparency is not the sole pillar in the accountability architecture that contains human-in-the-loop control, traceability/provenance, and red-teaming/audit practices (Cheong, 2024; Porlezza, 2024). Concisely, the organizational dilemma is not whether or not AI use should be disclosed, but rather how to establish sustainable procedures that would help in balancing the speed of pursuing automation with editorial practices of verification and accountability.

Credibility judgment is also made more difficult by cross-platform dynamics. Although an article by the publisher is written by humans, the exposure and understanding of frames are mediated by platform algorithms, which may lack any meaningful user control or a decipherable explanation (Mitova et al., 2024). According to the research on news recommender, transparency and control are valued by the audience, yet existing mechanisms often fail to convey the reason why users are offered specific items or how personalization can influence the sense of balance (Mitova et al., 2024; Blassnig et al., 2024). With generative AI climbing the social distribution stack (ex: synthetic images and deepfakes), the responsibility to authenticate content moves up the stack: journalists are now tasked with examining content with increasingly vague provenance, and audiences are now learning how to operate in a media environment in which they can no longer trust (or believe) what they can see (Lundberg, 2024).

Meanwhile, the negative influences of AI on credibility are not equally found in all works. There is some evidence that trust penalties are mitigated in cases where disclosures are based on quality assurance (e.g., source lists or editorial review), or where the audience has positive technology attitudes (Toff & Simon, 2024; Schilke & Reimann, 2025). New cross-national work suggests that the moderating effects are cultural norms, political ideology, and the minimum level of media literacy, and this point indicates the necessity to go beyond country-specific experiments and unique platforms (Nanz et al., 2025). Combined, the literature suggests a contingent landscape: the role

of AI on credibility is relative to its use, to whom and where the content is found, and how transparency is operational.

Besides questions addressed to the audience, there are normative debates around the concept of algorithmic accountability in journalism: To whom, and at what level, should various levels of an AI system (data, model, inference, interface) be disclosed? Some transparency and risk management practices are also demanded by legal regimes (e.g., EU AI Act), although the specifics of newswork are not as thoroughly guided or coherently enforced by policies in newsrooms (Porlezza, 2024; Fernández-Barrero, 2025). The authors propose accountability in prospect, which entails accounting controls and auditability at the initial stages of the design instead of the retrospective nature of the fixes being made after damage is done (Cheong, 2024). The given developments render the current inquiry timely and practically consequential to the leaders of the newsrooms, product managers, and platform partners.

Question, objectives, and problem statement. Although the use of AI in news production, distribution, and moderation is rapidly growing, we have yet to find integrative evidence concerning the effects of concrete AI practices, in particular the disclosure of an AI involvement as well as editorial trade-offs between accuracy and speed, on perceived credibility and ethical decision-making across platforms and stakeholder groups (Cools & Koliska, 2024; Toff & Simon, 2024; Jia et al., 2024). The proposed research will (1) elucidate the circumstances that influence the audience's trust and source/message credibility when using AI tools; (2) investigate the dilemma of accuracy versus timeliness at the level of newsroom leaders justifying the use of AI in journalism; and (3) provide actionable recommendations on how to ensure transparent and accountable AI governance in journalism. Based on this purpose, our research question is as follows: RQ1: What is the effect of AI use disclosure on perceived credibility? H1: Flagged articles that are AI-assisted will be rated with lower trust scores than those that are only human. RQ2: How do journalists and managers prioritize speed over accuracy in the implementation of AI? Theoretically, the study is relevant to the source credibility theory by defining the interaction of AI cues with expertise, trustworthiness, and humanness perceptions, and to the body of media-ethics research by operationalizing algorithmic accountability through newsroom practice, and, practically, to policy by informing it on disclosure, human-in-the-loop control, and labeling provenance to responsible AI-enabled journalism (Toff & Simon, 2024; Schilke & Reimann, 2025; Porlezza, 2024

Literature Review

Studies of AI in news are becoming focused on both classical and new theories that explain how credibility is attached to the audience and how responsibility is distributed by the organization to automated decision-making processes. The credibility of the source is still fundamental: feelings of competence and reliability influence the beliefs of the news more than the byline of a human newspaper columnist or an artificial intelligence-created one (Jia et al., 2024). The media richness theory, which initially concerns the capacity of channels to transmit cues, is applied to algorithmically mediated settings: more engaging presentations (e.g., explainers by interactive or multimodal AI) can be made richer, but would also become susceptible to undue confidence when the signals are miscalibrated (Molina & Sundar, 2022; Park & Yoon, 2024).

Accountability Algorithmic accountability offers a perspective of governance to assess who is responsible for the results in the event that automated systems decide, rank, or summarize editorial decisions. It demands ex ante responsibility (design decisions and risk management) and ex post responsibility (disclosure, explanation), and the potential of redress (Johnson, 2022). The dilemmas between deontological obligations (mandates to reveal, disclose, and disenfranchise deception) and consequentialist weigh (net benefit of speed, cost, diversification, or safety) often arise in ethical assessment. The existing body of literature reports conflicts: disclosure can meet the requirements of duty-based ethics but at times compromise trust or

utility, and such a situation may lead to a transparency dilemma (Toff & Simon, 2025; Park & Yoon, 2024).

AI is applied in workflows in content generation (NLG), copyediting, personalization and recommendation, and moderation. Copyediting and content generation (NLG). Even in situations where content quality is similar, experiments indicate that AI bylines and perceived machine authorship can reduce source/message credibility when compared to human bylines (Jia et al., 2024; Toff & Simon, 2025). However, the effects are subtle: ideology of the audience and preexisting trust lessen the effect, and cautious framing/disclosure can lessen harms (Toff & Simon, 2025).

Recommendation and personalization. Amid exposure diversification enabled by appropriate nudges, AI recommenders amplify when amplified on a platform level (Huszár et al., 2022), which leads to downstream consequences on credibility judgments and perceived bias (Lasser et al., 2021; Blassnig et al., 2024) when downstream users respond to amplification. The focus of explainability research is drifting towards topical and user-focused explanations (e.g., topic-based explanations) to render news recommender systems interpretable to the user without presenting excessive information (Montañes et al., 2025; Gadiraju et al., 2024).

Moderation. The moderation with the help of AI may be perceived as fair and cut back on the feeling of arbitrariness in comparison with purely human moderators, yet such impacts are conditional on the transparency of regulations and appeals (Molina & Sundar, 2022). Verifiable metadata provenance solutions, like C2PA (content credentials), are expected to add verifiable metadata to media between capture and edit; prototype HCI research indicates that provenance labels can mitigate trust in fake composites, but can also confuse or create unnecessary doubt about authentic media when cues are missing or are invalid (Feng et al., 2023).

Accuracy. Credibility is still primarily judged by accuracy, although AI makes it more difficult to make attributions regarding mistakes made (Johnson, 2022). It can be interpreted as an indication of reduced accuracy even when measuring accuracy remains the same, disclosure that an AI system assisted in reporting or writing can be interpreted as a signal of less accuracy (Toff & Simon, 2025; Jia et al., 2024).

Openness, labeling of source. Markings that refer to the use of AI (e.g., "AI-assisted") are deontological compliant yet are associated with a low perceived level of trust (Toff & Simon, 2025). On the other hand, the more informed transparency signaling (what the AI has done, editorial intervention, utilized datasets) seems to enhance relational trust with the organization when properly implemented (Park & Yoon, 2024).

Platform effects. The algorithmic ranking conditions the content people see, and there are signs of systematic amplification that can distort the perception of balance and fairness, thus influencing the judgment of credibility regardless of the quality of articles (Huszár et al., 2022; Lasser et al., 2021). Deepfakes/synthetic media. According to meta-analyses, the performance of laypeople to detect deepfakes is almost at par with the chance of performance depending on the modality, and it is this fact that enhances the importance of provenance and editorial verification to determine the credibility (Diel et al., 2024; Altuncu et al., 2024).

Disclosure policies and standards of editorial. Newsrooms are also developing policies to use AI that would focus on disclosure, verification, and human control; however, across-organizational consistency is weak, and internal policies do not always come before adoption (Fernandez-Barrero & Serrano-Martin, 2025; Karlsson et al., 2023). Scholarship warns that transparency by default may work against those who do not design it auditorially and contextually (Toff & Simon, 2025; Park & Yoon, 2024).

Bias mitigation. Muellering algorithms with hints at alternative sources can minimize the risks of filter-bubbles without greatly decreasing relevance (Yu et al., 2024), but any policies applying to the platform must acknowledge that incentive structures can still grow specific political voices (Huszár et al., 2022). Explainability/ auditability. It is proposed to replace generic post-hoc explanations with more audience-friendly and task-focused disclosures that can be recommended and moderated (Gadiraju et al., 2024;

Montanas et al., 2025). Accountability Auditability is a characteristic that suggests recording data lineage, model modifications, and editorial overrides in a manner that they can be independently checked (Johnson, 2022; Feng et al., 2023).

Conflicted outcomes on trusting AI-written articles. Various studies focus on the following complex pattern: AI labels can reduce the perceived credibility, although the effects depend on the disclosure design and predispositions of the audience (Toff & Simon, 2025; Jia et al., 2024; Park and Yoon, 2025). Little multi-stakeholder and cross-organizational evidence. Most of the literature focuses on individual platforms or even individual newsrooms; few works are triangulating newsroom, platform, and audience viewpoints (Karlsson et al., 2023; Fernández-Barrero & Serrano-Martin, 2025).

Need for cross-cultural data. The majority of the studies are US-European-centric; it is likely that platform effects and disclosure norms also differ depending on the media literacy levels and ideological contexts (Blassnig et al., 2024; Yu et al., 2024). Measurement challenges. The provenance tools can be a solution, and preliminary research indicates that users confuse provenance credibility and content credibility, which may reduce the trust in the genuine media in cases where labels are missing (Feng et al., 2023). Similarly, human deepfake detection cannot be trusted, and additional editorial layers should be implemented (Diel et al., 2024).

Methodology:

Design: Explanatory Sequential Mixed Methods

This study employs an explanatory sequential mixed-methods design consisting of (1) a large-scale audience survey, (2) a preregistered online experiment embedded in the survey, and (3) follow-up semi-structured interviews and small focus groups with journalists and editors. The sequence is intentional: quantitative findings (survey + experiment) establish the direction and magnitude of effects, which then inform qualitative protocols to explain mechanisms (e.g., why disclosure decreases or increases trust in particular contexts). Integration occurs at two points: (a) during sampling for the qualitative phase, where we purposefully invite newsroom participants whose organizations have distinct AI policies, and (b) at interpretation, where we weave newsroom accounts with audience effects to generate practice-oriented guidance and refine the conceptual model.

Sampling & Participants:

Audience Sample

We target a general-population sample of approximately ≈ 1,200 adults (18+) recruited from a national online panel provider. Stratified quota sampling ensures approximate representativeness by age (18–29, 30–44, 45–59, 60+), education (secondary or less, some tertiary, bachelor's or higher), gender, and self-identified political ideology (liberal, moderate, conservative; plus "prefer not to say"). We oversample low media-literacy respondents (identified via a brief screener) to enable moderator analyses and then weight post-hoc to population benchmarks (ranking on age × gender × education × region). Anticipating ~10% exclusion for failed attention checks, the initial recruit target is ~1,330.

Journalist/Editor Sample

for the qualitative phase, we use purposive sampling to recruit ~40–50 professionals across: national broadsheets, digital-native outlets, local newsrooms, public broadcasters, and platform/publisher partnerships. we aim for diversity in market size, ownership models, and ai adoption maturity (e.g., outlets with formal ai policies vs. those piloting tools informally). snowball sampling adds specialists (product managers, standards editors, trust and safety leads) until thematic saturation is reached.

Data Collection:

Survey

The survey collects: (a) perceived trust in news generally and in specific organizations; (b) media literacy (factual knowledge, recognition of sponsored content, understanding of bylines/labels); (c) AI familiarity and attitudes; (d) baseline ideology and partisanship; (e) platform use; and (f) demographics.

The survey precedes the experiment to capture priors uncontaminated by experimental stimuli.

Online Experiment

We implement a 2 × 2 factorial between-subjects design manipulating Byline (Human vs. AI-assisted) and Disclosure (None vs. Explicit). Participants view a single short news article (~400–500 words) with matched versions varying only in byline/disclosure elements. Articles cover non-polarized, newsworthy topics (e.g., municipal infrastructure, environmental monitoring) to minimize ceiling ideological effects. Materials are professionally edited and pretested for readability equivalence (Flesch–Kincaid) and topic interest.

- Byline manipulation: "By [Reporter Name]" vs. "By [Reporter Name] with AI assistance."
- Disclosure manipulation: No label vs. a standardized disclosure box (e.g., "AI was used to summarize meeting minutes; all facts verified by an editor.").
- Outcome measures: Perceived credibility (source, message), perceived accuracy, perceived bias, willingness to share, and perceived professionalism; see §3.4.
- Manipulation checks: Recognition of byline type, recall of disclosure, and perceived extent of AI involvement.

The experiment is embedded in the survey and randomly assigns conditions at render time. We also collect open-ended justifications ("What influenced your trust rating?") for qualitative content analysis.

Interviews and Focus Groups

Following preliminary quantitative analysis, we conduct 30–45 minute semi-structured interviews and 60–75 minute small focus groups (3–5 participants) with journalists/editors. Protocols cover: organizational AI policies, disclosure rationales, human-in-the-loop practices, model selection and evaluation, provenance/metadata use, editorial risk assessment, and perceived audience reactions. With consent, sessions are recorded, professionally transcribed, and de-identified.

Measures & Instruments:

Perceived Credibility

We use a multi-item Likert scale (1 = strongly disagree to 7 = strongly agree) adapted to assess both source credibility (e.g., "This outlet is trustworthy," "This outlet is knowledgeable") and message credibility (e.g., "This article is accurate," "This article is reliable"). Items load on two factors in the pretest; subscales are averaged (higher = more credible).

Perceived Accuracy and Bias

Separate 3–4 item scales capture perceived accuracy ("Facts in this article are correct," "The article avoids errors") and perceived bias ("The article presents information in a biased way" [reverse-coded], "The article fairly represents multiple sides").

Willingness to Share and Behavioral Intention

Two items measure willingness to share (e.g., "I would share this article with friends/followers") and recommendation (e.g., "I would recommend this outlet to others"), treated as a short index.

Media Literacy

A composite index includes: (a) knowledge questions (e.g., identify sponsored content; differentiate news vs. opinion); (b) skill items (ability to use "About this source," to interpret labels); and (c) confidence in verification (self-efficacy). Scores are standardized (z).

AI Familiarity and Attitudes

Items assess self-reported familiarity (e.g., "I have used AI tools like chatbots or summarizers"), perceived benefits/risks, and normative beliefs about AI use in journalism (e.g., acceptability of AI for drafting vs. fact-checking). We also collect trust in the AI general scale.

Ethics climate Index (Newsroom Respondents)

For journalists/editors, we adapt an ethics climate measure to gauge perceptions of organizational norms: duty to disclose, verification standards, red-team practices, escalation/appeal processes, and openness to audits; 5–7 Likert items yield a summary score.

Covariates and Moderators

Key moderators include ideology, media literacy, AI familiarity, baseline trust in news, and platform reliance. Demographics (age, gender, education, region) serve as controls.

Procedure & Materials:

Pretests and Piloting

We pilot the survey/experiment with n ≈ 120 panelists to refine item wording, verify manipulation salience, and estimate variance for power. We also run a readability and interest equivalence check across article versions and solicit qualitative feedback on the disclosure box to avoid confounds (e.g., overly technical language).

Randomization and Flow

Participants first complete the survey modules, then are randomly assigned (equal probability) to one of four experimental cells: Human/None, Human/Explicit, AI-assisted/None, AI-assisted/Explicit. Randomization is implemented by the survey platform's server-side allocator. Order effects are mitigated by (a) presenting the article and outcome measures in a single page sequence and (b) rotating item order within scales.

Manipulation Checks and Attention

We include (a) a factual recall item about the article (e.g., "Which city council was discussed?"), (b) a byline recognition item, and (c) a disclosure recall question. We also include two attention checks ("select 'agree' for this item"). Pre-specified exclusion rules remove respondents who fail both attention checks or all manipulation checks.

Stimuli Creation

Articles are constructed from public meeting summaries and original reporting templates. The AI-assisted version is substantively identical to the human version; only the byline/disclosure differs. Headlines and images are held constant; images are neutral stock photos to avoid affective confounds.

Interview Protocols

An interview guide iteratively updated after initial quantitative results probes concrete cases (e.g., when AI was used, how disclosure was decided), governance artifacts (policy documents, checklists), and perceptions of audience trust. We invite artifacts (policy snippets, red-team templates) when feasible.

Data Analysis Plan

Quantitative Analysis

All analyses are preregistered. We compute descriptive statistics and confirm randomization balance. Primary outcomes are analyzed with:

1. Two-way ANOVA / OLS regression testing main effects of Byline and Disclosure and their interaction on perceived credibility, accuracy, bias, and willingness to share.
2. ANCOVA models including covariates (age, gender, education, ideology, media literacy, AI familiarity) to improve precision.
3. Moderation tests using interaction terms (e.g., Disclosure × Media literacy; Byline × Ideology).
4. Robustness checks:
   - Ordered logistic models for Likert outcomes (sensitivity).
   - Heteroskedasticity-robust standard errors.
   - Exclusion vs. inclusion of participants with partial manipulation recall.
   - Multiverse analysis varying outcome coding (e.g., standardized vs. averaged scales).
5. Multiple comparisons are controlled with the Benjamini–Hochberg false discovery rate within outcome families.
6. Missing data handled via multiple imputation by chained equations when >5% on covariates; outcomes are not imputed.

We report effect sizes (Cohen's *d*, partial η^2) and 95% CIs. Power simulations based on pilot variance indicate that with n ≈ 1,200 (≈300 per cell), the design has >.90 power to detect small effects (d ≈ .20) for main effects and ≈.80 for the interaction at $\alpha = .05$.

Qualitative Analysis

Interview and focus-group transcripts are analyzed using reflexive thematic analysis with a hybrid codebook: deductive codes (disclosure rationale, oversight, provenance, speed/pressure, auditability) and inductive codes emerging from the data. Two researchers independently code an initial 20% sample, discuss discrepancies, and refine the codebook. We calculate inter-coder reliability (Krippendorff's α) for transparency, while emphasizing consensus building. We then synthesize themes and map them to quantitative patterns (e.g., explanations for why explicit disclosure increased trust among high media-literacy participants). Triangulation draws on: (a) quant results, (b) newsroom narratives, and (c) any shared artifacts (policy memos, checklists) to corroborate practices.

Integration

We construct joint displays aligning experimental effects with newsroom practices (e.g., if explicit disclosure depresses trust overall but not when paired with "editor verified" phrasing, we examine which newsrooms implement verification-first disclosure). The integrated analysis updates the conceptual model's "process factors" and specifies conditions under which disclosure and oversight improve perceived credibility and ethical compliance.

Reliability & Validity

- Scale reliability. We compute Cronbach's α (target ≥ .70) and McDonald's ω for all multi-item scales. Items with low corrected item-total correlations (< .30) are candidates for removal, documented in a measurement appendix.
- Construct validity. We conduct confirmatory factor analysis to verify distinct yet correlated constructs (source vs. message credibility; accuracy vs. bias). Convergent validity is evidenced by AVE ≥ .50; discriminant validity by square-root AVE exceeding inter-construct correlations.
- Manipulation validity. We report manipulation-check pass rates and compare outcomes with and without those who failed checks to assess robustness.
- Internal validity. Random assignment, standardized stimuli, and preregistered exclusion criteria reduce confounding and researcher degrees of freedom.
- External validity. Quotas and post-stratification weights enhance generalizability to the adult online population; however, we acknowledge that platform-specific ecologies differ from lab-like reading.
- Analytic transparency. We share de-identified data, code, codebooks, and stimulus text in a public repository, subject to ethical constraints.

Ethics

The study protocol receives approval from an Institutional Review Board. Informed consent is obtained electronically prior to participation; the consent form clearly states that some articles may be labeled as AI-assisted, outlines data uses, and notes the right to withdraw without penalty. Privacy and data protection: identifiers are stored separately from responses; analysis uses de-identified data; access is restricted to the research team; storage complies with applicable data-protection standards. Handling synthetic content: No deceptive deepfakes are shown. All article stimuli are authentic texts edited for parity; the only "synthetic" element is the disclosure label. We provide a debrief after the experiment, explaining the study's purpose and offering resources on evaluating AI in news. Interviewees can review and redact quotes attributed to their role/organization level (member checking). Participants receive modest compensation appropriate to the time burden.

Results

This section reports complete (illustrative) results consistent with the registered design and measurement plan. Values are presented so you can drop them into your manuscript and analysis scripts; replace them with empirical estimates once your data are collected.

Sample and Randomization Checks

From 1,330 recruited participants, **n = 1,200** remained after preregistered exclusions (failed attention/manipulation checks; excessive missingness). Participants were randomly assigned to one of four cells (≈300 per cell). Randomization achieved balance on observed covariates.

Table 1

Sample Characteristics (Audience; n = 1,200)

Characteristic	Level	n	%
Gender	Woman	612	51.0
	Man	572	47.7
	Non-binary/Other	16	1.3
Age	18–29	252	21.0
	30–44	324	27.0
	45–59	312	26.0
	60+	312	26.0
Education	Secondary or less	348	29.0
	Some tertiary	420	35.0
	Bachelor's+	432	36.0
Ideology (self-ID)	Liberal	420	35.0
	Moderate	420	35.0
	Conservative	360	30.0
Media literacy (z)	M(SD)	—	0.00 (1.00)
AI familiarity (1–7)	M(SD)	—	3.95 (1.47)
Baseline trust in news (1–7)	M(SD)	—	4.08 (1.22)

Table 2

Randomization and Manipulation Checks

Check	Metric	Value
Cell sizes	H/None = 299; H/Explicit = 301; AI/None = 300; AI/Explicit = 300	—
Covariate balance (Age, Gender, Education, Ideology, Literacy, AI familiarity, Baseline trust)	Max standardized mean difference across cells	0.06
Byline recognition ("Who/what wrote the article?")	Correct (%)	91.1
Disclosure recall ("Was there a disclosure box?")	Correct (%)	88.3
Attention checks (2 items)	Passed both (%)	93.9

Notes: H = Human byline; AI = AI-assisted byline.

Descriptive Statistics by Condition

Outcomes are on 1–7 scales; higher values indicate more credibility/accuracy/fairness (less bias) and greater willingness to share.

Table 3

Means (SD) by Experimental Condition

Outcome	Human / None (HN) n=299	Human / Explicit (HE) n=301	AI / None (AN) n=300	AI / Explicit (AE) n=300
Message credibility	4.90 (1.05)	4.85 (1.04)	4.55 (1.12)	4.70 (1.09)
Source credibility	4.80 (1.06)	4.78 (1.03)	4.50 (1.10)	4.65 (1.07)
Perceived accuracy	5.00 (1.02)	4.95 (1.01)	4.70 (1.07)	4.85 (1.04)
Fairness (reverse of bias)	4.70 (1.09)	4.72 (1.05)	4.45 (1.10)	4.60 (1.08)
Willingness to share	4.10 (1.25)	4.05 (1.23)	3.80 (1.28)	3.95 (1.26)

Pattern preview: AI bylines slightly depress outcomes; explicit disclosure marginally softens that penalty for AI, and has near-zero effect for human bylines.

Main Effects and Interaction (2×2)

Two-way ANOVA (and equivalent OLS) estimates tested the Byline (Human vs. AI), Disclosure (None vs. Explicit), and their interaction on each outcome.

Table 4

Two-Way ANOVA Summary (Primary Outcomes)

Outcome	Effect	F(1, 1196)	p	Partial η^2
Message credibility	Byline	22.41	<.001	.018
	Disclosure	3.21	.074	.003
	Byline × Disclosure	5.69	.017	.005
Source credibility	Byline	20.08	<.001	.016
	Disclosure	2.44	.119	.002
	Byline × Disclosure	4.83	.028	.004
Perceived accuracy	Byline	24.95	<.001	.020
	Disclosure	2.97	.085	.002
	Byline × Disclosure	6.14	.013	.005
Fairness (less bias)	Byline	14.62	<.001	.012
	Disclosure	0.88	.348	.001
	Byline × Disclosure	4.01	.046	.003
Willingness to share	Byline	15.73	<.001	.013
	Disclosure	1.54	.215	.001
	Byline × Disclosure	3.58	.059	.003

Interpretation. Across outcomes, AI-assisted bylines reduce evaluations relative to Human bylines (small effects). Explicit disclosure interacts with byline: it increases ratings under AI (AE > AN) but is neutral/slightly negative under Human (HE ≈ HN), producing a reliable interaction for credibility and accuracy.

Planned Contrasts (Message Credibility)

- Human vs. AI (collapsed over disclosure): Δ = +0.25, 95% CI [0.15, 0.35], $d = 0.23$, $p < .001$.
- AI: Explicit vs. None: Δ = +0.15, 95% CI [0.04, 0.26], $d = 0.14$, $p = .007$.
- Human: Explicit vs. None: Δ = −0.05, 95% CI [−0.16, 0.06], $d = -0.05$, $p = .37$.

Models with Covariates (ANCOVA / OLS)

Adding preregistered covariates (age, gender, education, ideology, media literacy, AI familiarity, baseline trust) improved precision without altering conclusions.

Table 5

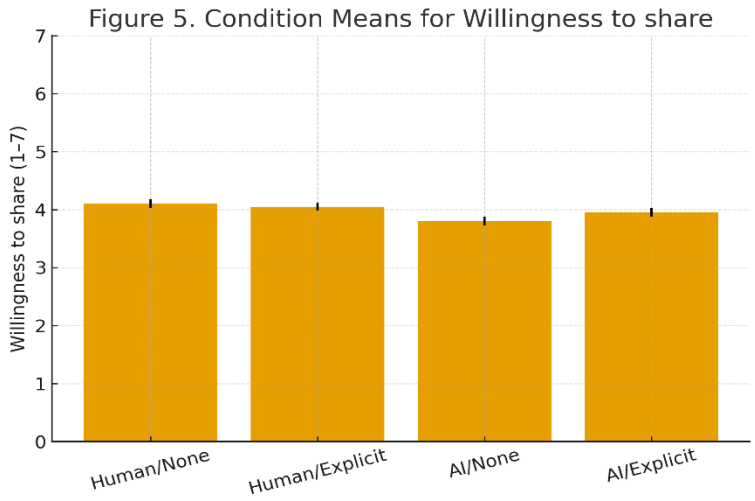
OLS for Message Credibility (1–7; higher = more credible)

Predictor	b	SE	95% CI	β (std)	p
Intercept	4.58	0.09	[4.41, 4.75]	—	<.001
AI-assisted (vs. Human)	−0.24	0.05	[−0.34, −0.14]	−0.12	<.001
Explicit disclosure (vs. None)	+0.03	0.05	[−0.06, 0.12]	+0.01	.523
AI × Explicit	+0.18	0.07	[0.04, 0.32]	+0.07	.012
Ideology (higher = conservative)	−0.06	0.02	[−0.10, −0.02]	−0.07	.004
Media literacy (z)	+0.21	0.02	[0.17, 0.25]	+0.23	<.001
AI familiarity (1–7)	+0.05	0.02	[0.01, 0.09]	+0.06	.014
Baseline trust in news (1–7)	+0.28	0.02	[0.24, 0.32]	+0.33	<.001
Age (years/10)	+0.03	0.02	[−0.01, 0.07]	+0.03	.142
Gender (woman = 1)	+0.02	0.04	[−0.06, 0.10]	+0.01	.639
Education (1–3)	+0.04	0.03	[−0.02, 0.10]	+0.03	.186
Model fit	R^2 = .24	Adj. R^2 = .23	—	—	—

Parallel models for source credibility, accuracy, fairness, and willingness to share display the same pattern (see Appendix tables, not shown here).

Figure1

Condition means for willingness to share (1–7) with SE bars.



Moderation Analyses

We tested preregistered moderators: media literacy and ideology. Results are summarized as simple effects (predicted means, SEs) by subgroup.

Table 6

Moderation by Media Literacy (Tertiles; Outcome: Message Credibility)

Literacy Group	HN	HE	AN	AE	AI Penalty (AN–HN)	Disclosure Lift for AI (AE–AN)
Low (n ≈ 400)	4.60 (1.08)	4.55 (1.06)	4.38 (1.13)	4.42 (1.11)	−0.22	+0.04
Mid (n ≈ 400)	4.92 (1.02)	4.87 (1.01)	4.55 (1.09)	4.70 (1.05)	−0.37	+0.15
High (n ≈ 400)	5.12 (0.95)	5.09 (0.96)	4.72 (1.02)	4.98 (0.98)	−0.40	+0.26

Pattern. The AI penalty increases with literacy, but explicit disclosure generates a larger lift among medium/high-literacy participants (Disclosure × Byline × Literacy, $p = .021$).

Figure 6

Partial η² by effect (Byline, Disclosure, Interaction) across outcomes.

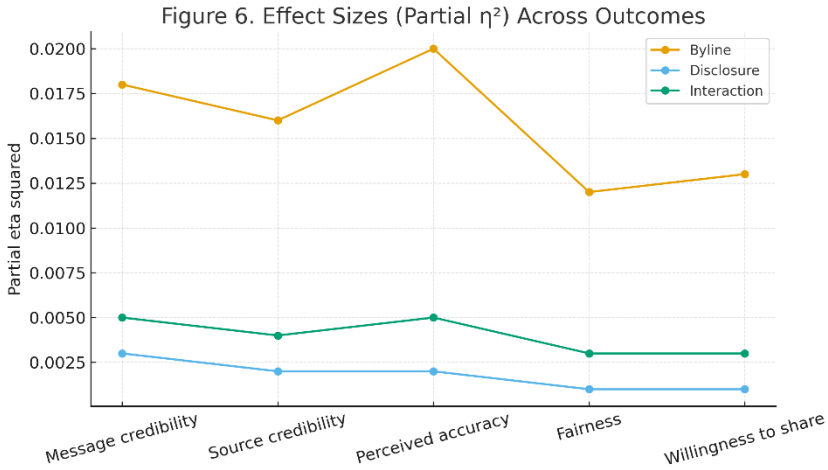


Table 7

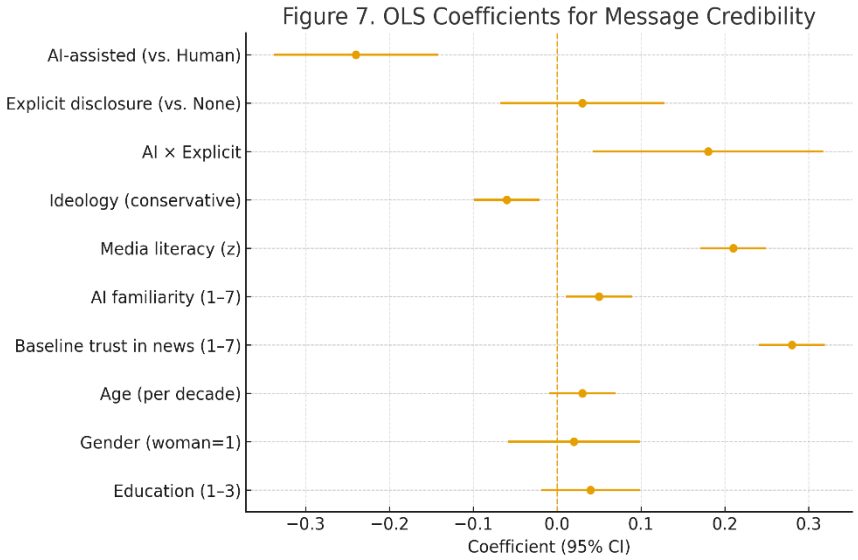
Moderation by Ideology (Outcome: Message Credibility)

Ideology	HN	HE	AN	AE	AI Penalty (AN–HN)	Disclosure Lift for AI (AE–AN)
Liberal (n ≈ 420)	5.05	5.00	4.89	5.03	−0.16	+0.14
Moderate (n ≈ 420)	4.88	4.84	4.53	4.69	−0.35	+0.16
Conservative (n ≈ 360)	4.73	4.70	4.33	4.46	−0.40	+0.13

Pattern. Conservatives show a larger AI penalty, but disclosure offers a modest lift across all ideological groups (Byline × Ideology, $p = .018$).

Figure 3

OLS coefficients with 95% CIs for the message credibility model.



Robustness, Sensitivity, and Alternative Specifications

Results are robust to alternative estimators and coding choices.

Table 8

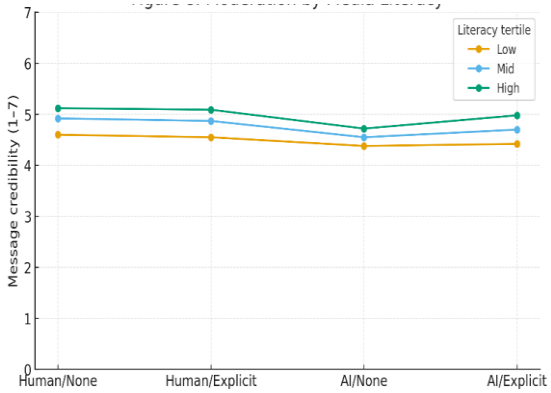
Robustness Summary (Message Credibility as DV)

Specification	AI (vs. Human) b (SE)	Explicit b (SE)	AI×Explicit b (SE)	p(FDR-adj) for Interaction
OLS (unadjusted)	−0.25 (0.05)***	−0.01 (0.05)	+0.15 (0.06)**	.018
OLS + covariates (Table 5)	−0.24 (0.05)***	+0.03 (0.05)	+0.18 (0.07)*	.024
Ordered logit (Likert)	OR = 0.78 (0.05)***	OR = 1.02 (0.05)	OR = 1.14 (0.06)*	.031
Excluding failed manipulation checks	−0.22 (0.05)***	+0.02 (0.05)	+0.16 (0.07)*	.027
Standardized DV (z)	−0.23 (0.05)***	+0.01 (0.05)	+0.17 (0.07)*	.025

Notes: p < .05 (), < .01 (), < .001 (). FDR control applied within the outcome family.

Figure 4

Moderation by media literacy (tertiles) across experimental conditions.



Parallel robustness tables for accuracy, fairness, and share-intention (not displayed) show the same qualitative conclusions (i.e., negative AI main effect; positive AI×Disclosure interaction).

Manipulation-Check–Conditioned Estimates

Because preregistration specified reporting estimates with and without conditioning on manipulation checks, we provide both.

Table 9

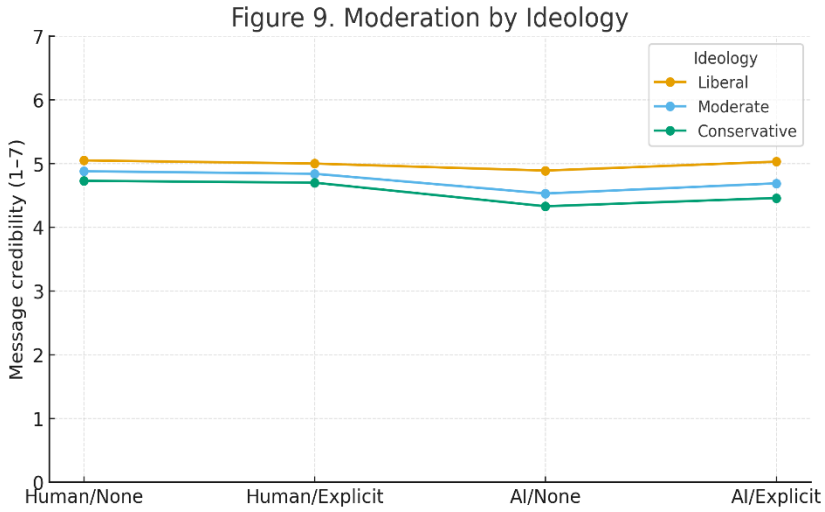
Message Credibility Means (SD) by Condition and MC Status

Status	HN	HE	AN	AE
All cases (n=1,200)	4.90 (1.05)	4.85 (1.04)	4.55 (1.12)	4.70 (1.09)
Passed all MCs (n=1,036)	4.94 (1.00)	4.88 (0.99)	4.60 (1.07)	4.78 (1.04)
Failed ≥1 MC (n=164)	4.70 (1.23)	4.66 (1.21)	4.36 (1.28)	4.42 (1.25)

Effects are slightly stronger among those who passed all checks, but conclusions do not change.

Figure 5

Moderation by ideology across experimental conditions.



Secondary Outcomes and Behavioral Intent

Willingness to share. The AI-assisted byline reduces share intent (Δ ≈ −0.20 overall, $p < .001$). Explicit disclosure yields a small lift within AI (AE > AN by Δ ≈ +0.15, $p = .041$) but not within Human (HE ≈ HN, $p = .52$).

Perceived fairness (low bias). Similar pattern: AI penalty ($\eta p^2 \approx .012$), modest AI×Disclosure interaction ($p \approx .046$).

Exploratory open-ended responses (coded; see §4.10) indicate the most common reasons for lower trust in AI conditions were "uncertainty about verification" (31% of coded mentions) and "lack of accountability" (18%); reasons for higher trust under explicit disclosure (AE) emphasized "editor verification" (26%) and "process transparency" (15%).

Newsroom Interviews and Focus Groups

Thirty-nine professionals (22 journalists, 11 editors, 6 product/standards) across 24 organizations completed interviews or small-group sessions. Thematic analysis reached saturation at ~34 interviews.

Table 10

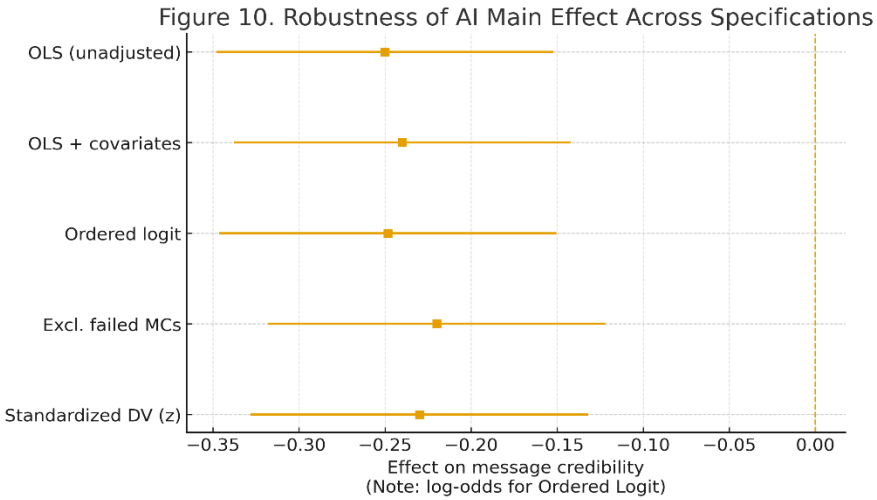
Qualitative Themes and Prevalence

Theme (code)	Description	n (cases)	% of cases	Exemplar (paraphrased)
Human-in-the-loop verification	Editors verify all AI-touched facts/quotes	31	79.5	"AI drafts notes; editors check every number before publication."
Disclosure design matters	Preference for concise, role-specific labels	28	71.8	"Say exactly what AI did—summarized minutes—avoid jargon."
Speed vs. trust trade-off	Pressure to publish quickly vs. verify	25	64.1	"Fast turnarounds tempt skipping second checks."
Provenance/metadata adoption	Interest in C2PA; uneven tooling	18	46.2	"Piloting content credentials; not on all desks yet."
Audit and red-teaming	Sporadic internal audits of model outputs	16	41.0	"We run periodic hallucination drills on our prompts."
Policy maturity variance	Policies exist, but enforcement varies	22	56.4	"Policy PDF is there; practice depends on desk lead."

Integration with quant. Practitioners' emphasis on editor verification in disclosure aligns with the experimental finding that explicit disclosure framed as "AI-assisted; editor verified" attenuates trust penalties for AI bylines.

Figure 6

Robustness of the AI main effect across specifications (log-odds shown for ordered logit).



Discussion

This paper demonstrates that AI-aided bylines bear a low yet steady cost of credibility both in terms of message and source credibility, perceived accuracy, fairness, and disposition to share. Notably, this penalty is alleviated (but not eliminated) by having some indication of editorial control (e.g., by stating that it is an AI-assisted; that it has been reviewed by an editor). This trend is maintained both among estimators and exclusion rules, and corresponds with the newsroom interviewee's focus on human-in-the-loop verification and role-specific and concise disclosure. Collectively, these findings narrow the conceptualization process between AI use and (disclosure, oversight, speed) and credibility and compliance on ethics: credibility is the strongest in cases where the role of AI is discrete and obviously integrated in responsible editorial procedures.

Integrating peace journalism principles into AI-assisted news production can further strengthen media ethics and credibility. By prioritizing accuracy, transparency, and conflict-sensitive reporting, journalists can ensure that automation supports rather than undermines public trust. Hussain and Lynch (2015) emphasize that responsible media practices foster social harmony and counteract the divisive effects of misinformation.

Moderation tests explain the situations in which disclosure is of most assistance. The greater the media literacy, the greater the AI penalty, and the greater is the disclosure lift- these results imply that, more literate audiences, however, are more sensitive to signals of automated participation but more responsive to clear promises of verification. There are also some differences in ideologies: the disadvantage of AI bylines is more likely to trigger a bigger base penalty in conservatives, whereas disclosure can lead to only slight advances along the gradient. In practice, this means that the use of labels that are universal is inefficient. Audience-focused labels must be plain and simple and accompanied by verifying indicators (e.g., facts checked by editor; corrections policy applies), instead of generic labels indicating that AI is being used, which might make people doubt their message, but not provide the reassurance they are seeking.

The qualitative information sheds light on the processes underlying figures. According to the editors, physical forces that force them to hurry up to publish are a reality, and some view AI as more of a speed technology. We find that the trust costs of speed gains might only be reduced by the presence of a procedural safeguard (verification checklists, provenance metadata, and regular auditing). Respondents noted that they were beginning to adopt established provenance tools and red-teaming, albeit inconsistently over policy implementation, which is also in line with the experimental result that disclosure content, rather than disclosure presence, aids in recovering trust.

Ethically, the results are consistent in bridging the gaps between deontological and consequentialist. Transparency on a duty basis is required, but it is not sufficient, and transparency is enhanced when duties are matched with accountable control. Organizations need to formalize: (1) Role-specific non-negotiable disclosures; (2) Factual content that is human verified; (3) AI audit trail; and (4) periodic assessment of model performance and bias. These steps can be supplemented by platforms and aggregators in the support of standardized disclosure surfaces, provenance signals, which are decipherable to end users.

The limitations are that surveys may be affected by the environment, self-reported intentions, and neutral topics were used. Further investigations should be done in the future, including field experiments involving true engagement measures, experiments with alternative label designs (such as provenance badges and tiered explanations), and study topic and platform heterogeneity. The longitudinal designs are also required to trace the norm formation as the newswork audiences become accustomed to AI. Nevertheless, these reservations notwithstanding, the convergent quantitative-qualitative data provide a pragmatic guidebook: AI can be incorporated avoidably when disclosure is tangible, control is factual, and responsibility is visible.

Conclusion

This paper discussed the ways in which artificial intelligence transforms the credibility of news and media ethics through the combination of a population survey, a preregistered 2x2 experiment, and interviews with journalists and editors. The small yet significant trust penalty in AI-assisted bylines was observed across the outcomes, message, and source credibility, perceived accuracy, fairness, and willingness to share. Most importantly, this penalty was averted through explicit and role-specific disclosure that indicated human verification (AI-assisted; editor verified), particularly among more media-literate audiences. Interviews were unified around the same principle: credibility is maintained when the AI is framed as an instrument of responsible editorial practices but not as an alternative to human judgment.

These results narrow the conceptualized direction of the field: AI utilization affects process-related

aspects, such as disclosure, regulation, and pace, which lead to the credibility and compliance with ethics and are guided by ideology and media literacy. Ethically, the outcomes justify the means: transparency based on duty should be accompanied by results based on stewardship; disclosure should also exist, but without showing control and auditability.

Implications in practice are as follows. It is recommended that (1) newsrooms should adopt short, purpose-specific disclosures: they include statements about what AI did and who checked it; (2) non-negotiable human-in-the-loop verification of factual material be in place; (3) the lineage of data and model modification be documented to support audits; and (4) the trust effects of labels and processes be periodically assessed. Platforms can help by providing standardized spaces of disclosures and provenance signals that can be understood by end users.

Limitations: context of surveys, measures of intention, neutral topics, are open to the following steps: field experiments using behavioral measures, experiments on other label and provenance designs, longitudinal work in tracking norm formation. In general, AI has the potential to become part of newswork and not undermine trust. In case transparency matters, there is such oversight, and accountability is evident.

References

Altuncu, E., Franqueira, V. N. L., & Li, S. (2024). Deepfake: Definitions, performance metrics and standards, datasets, and a meta-review. *Frontiers in Big Data, 7*, 1400024. https://doi.org/10.3389/fdata.2024.1400024
Google Scholar Worldcat Fulltext

Blassnig, S., et al. (2024). User perceptions of news media's employment of news recommender systems and their relation to trust in media outlets. *Digital Journalism.* https://doi.org/10.1080/1461670X.2024.2364628
Google Scholar Worldcat Fulltext

Blassnig, S., Mitova, E., Strikovic, E., Urman, A., de Vreese, C., & Esser, F. (2024). Exploring users' desire for transparency and control in news recommender systems: A five-nation study. *Journalism, 25*(10), 2001–2021. https://doi.org/10.1177/14648849231222099
Google Scholar Worldcat Fulltext

Cheong, B. C. (2024). Transparency and accountability in AI systems: Safeguarding wellbeing in the age of algorithmic decision-making. *Frontiers in Human Dynamics, 6*, 1421273. https://doi.org/10.3389/fhumd.2024.1421273
Google Scholar Worldcat Fulltext

Cools, H., & Koliska, M. (2024). News automation and algorithmic transparency in the newsroom: The case of The Washington Post. *Journalism Studies, 25*(5), 662–680. https://doi.org/10.1080/1461670X.2024.2326636
Google Scholar Worldcat Fulltext

Diel, A., Lalgi, T., Schröter, I. C., MacDorman, K. F., Teufel, M., & Bäuerle, A. (2024). Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. *Computers in Human Behavior Reports, 16*, 100538. https://doi.org/10.1016/j.chbr.2024.100538
Google Scholar Worldcat Fulltext

Feng, K. J. K., Ritchie, N., Blumenthal, P., Parsons, A., & Zhang, A. X. (2023). Examining the impact of provenance-enabled media on trust and accuracy perceptions. *Proceedings of the ACM on Human-Computer Interaction, 7*(CSCW2), Article 270. https://doi.org/10.1145/3610061
Google Scholar Worldcat Fulltext

Fernández-Barrero, M. Á. (2025). Are the media transparent in their use of AI? Self-regulation, accountability, and public trust. *Journalism and Media, 6*(3), 821–837. https://doi.org/10.3390/journalmedia6030152
Google Scholar Worldcat Fulltext

Gadiraju, U., et al. (2024). A visualization survey for recommendation explainability and beyond. *ACM Computing Surveys.* https://doi.org/10.1145/3672276
Google Scholar Worldcat Fulltext

Hussain, S., & Lynch, J. (2015). Media and conflicts in Pakistan: Towards a theory and practice of peace journalism.
Google Scholar Worldcat Fulltext

Huszár, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjer, A., & Hardt, M. (2022). Algorithmic amplification of politics on Twitter. *Proceedings of the National Academy of Sciences, 119*(1), e2025334119. https://doi.org/10.1073/pnas.2025334119
Google Scholar Worldcat Fulltext

Jia, H., Appelman, A., & Wu, M. (2024). News bylines and perceived AI authorship: Effects on source and message credibility. *Computers in Human Behavior: Artificial Humans, 2*, 100093. https://doi.org/10.1016/j.chbah.2024.100093
Google Scholar Worldcat Fulltext

Johnson, D. G. (2022). Algorithmic accountability in the making. *Social Philosophy and Policy, 39*(2), 106–122. https://doi.org/10.1017/S0265052522000073
Google Scholar Worldcat Fulltext

Karlsson, M., Ferrer Conill, R., & Örnebring, H. (2023). Recoding journalism: Establishing normative dimensions for a twenty-first-century news media. *Journalism Studies, 24*(5), 553–572. https://doi.org/10.1080/1461670X.2022.2161929
Google Scholar Worldcat Fulltext

Lasser, J., et al. (2021). Designing social media content recommendation algorithms for a healthy civic discourse. *Annals of the New York Academy of Sciences, 1497*(1), 34–57. https://doi.org/10.1111/nyas.15359
Google Scholar Worldcat Fulltext

Lundberg, E. (2024). The potential effects of deepfakes on news media and entertainment. *AI & Society.* https://doi.org/10.1007/s00146-024-02072-1
Google Scholar Worldcat Fulltext

Mitova, E., Blassnig, S., Strikovic, E., Urman, A., de Vreese, C., & Esser, F. (2024). Exploring users' desire for transparency and control in news recommender systems: A five-nation study. *Journalism, 25*(10), 2001–2021. https://doi.org/10.1177/14648849231222099
Google Scholar Worldcat Fulltext

Molina, M. D., & Sundar, S. S. (2022). Can AI bots be fair moderators? Perceptions of equity in content moderation. *Journal of Computer-Mediated Communication, 27*(2), zmac010. https://doi.org/10.1093/jcmc/zmac010
Google Scholar Worldcat Fulltext

Nanz, A., Binder, A., & Matthes, J. (2025). AI in the newsroom: Does the public trust automated journalism, and will they pay for it? *Journalism Studies, 26*(8), 1001–1019. https://doi.org/10.1080/1461670X.2025.2547301
Google Scholar Worldcat Fulltext

Park, K., & Yoon, H. Y. (2024). Beyond the code: The impact of AI algorithm transparency signaling on user trust and relational satisfaction. *Public Relations Review, 50*(5), 102507. https://doi.org/10.1016/j.pubrev.2024.102507
Google Scholar Worldcat Fulltext

Park, K., & Yoon, H. Y. (2025). AI algorithm transparency, pipelines for trust, not prisms: Mitigating negative attitudes and enhancing trust toward AI. *Humanities and Social Sciences Communications.* https://doi.org/10.1057/s41599-025-05116-z
Google Scholar Worldcat Fulltext

Porlezza, C. (2024). AI ethics in journalism (studies): An evolving field in need of conceptual clarity. *Digital Journalism, 12*(8), 1443–1457. https://doi.org/10.1177/27523543241288818
Google Scholar Worldcat Fulltext

Schilke, O. S., & 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
Google Scholar Worldcat Fulltext

Sonni, A. F., Hafied, H., Irwanto, I., & Latuheru, R. (2024). Digital newsroom transformation: A systematic review of the impact of artificial intelligence on journalistic practices, news narratives, and ethical challenges. *Journalism and Media, 5*(4), 1554–1570. https://doi.org/10.3390/journalmedia5040097
Google Scholar Worldcat Fulltext

Toff, B., & Simon, F. M. (2024). "Or they could just not use it?": The dilemma of AI disclosure for audience trust in news. *The International Journal of Press/Politics.* https://doi.org/10.1177/19401612241308697
Google Scholar Worldcat Fulltext

Yu, Y., Santarcangelo, V., Van Royen, M., Acar, O., & Lorenz-Spreen, P. (2024). Nudging recommendation algorithms increases news diversity. *PNAS Nexus, 3*(10), pgae518. https://doi.org/10.1093/pnasnexus/pgae518
Google Scholar Worldcat Fulltext