Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146
Vol. 3 No. 11 (November) (2025)

Disclosure of AI-Use in News Production: Effects on Audience Trust and News Credibility

Dr. Aniqa Ali
Assistant Professor, Department of Media and Communication Studies, International Islamic University, Islamabad. Email: aniqa.naseer@iiu.edu.pk

Dr. Amrat Haq
Assistant Professor, Department of Media and Communication Studies, International Islamic University, Islamabad. Email: amrat.haq@iiu.edu.pk

Dr. Adiba Akhtar
Lecturer, Department of Media and Communication Studies, National University of Modern Languages, Islamabad. Email: abakhtar@numl.edu.pk

Abstract
The utilization of AI in newsrooms is growing, and it is becoming a question of whether and how this information should be disclosed. The preregistered between-subjects online experiment (N=500) we report involved Disclosure (none vs disclosed) and Task-Type (assistive vs generative), exploratory variations in wording (assisted by AI vs written by AI), and placement variations (byline vs endnote). Article trust and news credibility were the primary results, perceived transparency was a mediator, and preregistered moderators were AI familiarity, media skepticism, political ideology, and topic involvement. Disclosure was reliable in enhancing transparency (a-path) but had a small negative impact on trust and credibility, and larger penalties in case AI was made to sound generative. The mediation showed both positive indirect with transparency and negative direct effects of disclosure with small net decreases. First stage moderation demonstrated greater transparency gains among AI-familiar viewers and lesser gains among media-skeptical viewers, second stage moderation demonstrated transparency less translated to trust in right-leaning respondents and less translated to credibility under high involvement. Wording aided was better than written, and the location of endnotes was safer overall.
Keywords: AI disclosure; transparency; trust; news credibility; automated journalism; algorithm aversion; wording/labeling; human–AI collaboration; media skepticism; political ideology.

INTRODUCTION

The rate at which newsrooms are integrating generative artificial intelligence (AI) into their workflows is astonishing, as it implements tools to generate ideas, backgrounding, summarization, transcription, headline testing, audience targeting and even full-text drafting. Although advocates view these systems as multipliers of forces that can liberate journalists to avoid tedious workloads and innovate, viewers are still weighing the way to interpret machine intervention in the news they watch. According to recent experiments, the mere mention of the fact that AI was helpful can reduce the perceived accuracy and desire to share even in cases when the content is correct, due to the perception of full automation and lack of supervision (Altay and Gilardi, 2024). Simultaneously, policy debate and platform movements demand naming and reporting in order

ISSN Online: 3007-3154
ISSN Print: 3007-3146
Vol. 3 No. 11 (November) (2025)

to restrict harms caused by synthetic media. The convergence of rapid adoption, competing intuitions of the audience and new disclosure regimes poses a high-stakes, timely question to journalism, when, where, and how should news organizations report about AI use without compromising credibility?

The transparency ideal has been preached since ancient times as the means to strengthen trust: demonstrate your work, clarify your techniques, and credibility is due to follow. That logic appears to be obvious in the context of AI but the emerging evidence is mixed. Cues regarding automation have been found to decrease perceived credibility, or create a form of AI aversion even with accurate content; it has also been found to have subtle or topic-dependent effects, or has been found to provide useful information about processes without significant trust costs. E.g. the results of recent large-N online experiments on the comparison of various label framings to AI-generated media indicate that process-related and harm-related labeling have different trade-offs in understanding, distrust and downstream behavior (Wittenberg et al., 2025). Byline and authorship indicators (written by staff with AI support and written by AI) can have a quantifiable effect on source and message credibility through perceived humanness in journalism-related scenarios (Jia et al., 2024). Combining these results, it can be implied that disclosure design, rather than the existence of a label, can be critical.

The second tension is that transparency may have a counterproductive effect when it puts in the limelight aspects that the audiences do not trust. The effect of algorithm aversion Research indicated that human judgment is frequently better than an algorithm, particularly when algorithms demonstrate errors or when the activity is subjective (Mahmud et al., 2022). A review of integrative style states that the phenomenon is not unitary: there are two aspects of aversion and appreciation, which are enabled by personal traits (e.g., literacy, trust propensity), characteristics of a task (e.g., evaluative vs. mechanical) and perceived responsibility (Jussupow et al., 2024). Applying this to newswork would mean that the cost of disclosing AI as a core author might be higher than disclosing AI as an assistive technology (e.g., transcription or translation) especially to audiences with a lower level of familiarity with AI. That reasoning is consistent with the indicators of discovery offered by observational and case-study research by news organizations that are struggling with algorithmic disclosure and internal accountability- where the news outlets are enjoying earnings of legitimacy in terms of transparency, but worry about being misunderstood by their audience or face reputational harm (Cools and Koliska, 2024, Opdahl et al., 2023).

Simultaneously, disclosure is not merely reputational risk management, but it is part and parcel of responsible AI deployment. The AI presence labelling is one of the major pillars of the platform and policy reaction to the synthetic content and misinformation that are supposed to aid users in shaping proper skepticism and preventing harm. However, even in a non-hard-news situation, the labels have the ability to make perceived accuracy and sharing decline without causing viewers to equate AI-generated with false (Altay and Gilardi, 2024). This means that newsroom disclosure policy needs to strike a balance between two goods, which at times work in opposite directions; (1) respect to transparency norms and audience autonomy, and (2) negative credibility punishment with no gratuity when human editorial control is strong. A better understanding of how disclosure can inform and not alarm and how wording and

ISSN Online: 3007-3154
ISSN Print: 3007-3146
Vol. 3 No. 11 (November) (2025)

positioning can reduce the spillovers associated with algorithm-aversion has immediate utility in practical terms when it comes to editors.

Most importantly, the issue with disclosure that journalism is supposed to take is not whether to label or not, but what exactly to say and where. Such transparency of process (AI used to transcription and translation; the story was written and edited by a human) plausibly has a different feel than authorship attribution (co-written by AI), and placement (front-of-article versus endnote versus expandable How we reported this box) likely has a salience and interpretation effect. Moderate average effects with heterogeneity under user predispositions and type of content (warnings about universality) of your news-labeling have been found in experimental work in neighboring news-labeling fields (e.g., credibility and opinion labels), which are not supposed to have those (Waddell, 2025; Aslett et al., 2022). Political situation is also relevant: recent findings on AI-attributed news reveal that partisan orientations and polarization of topics can precondition the role of attribution to create selection and credibility decisions (Zoizner et al., 2025). A research design where the words of disclosure and extent of tasks are manipulated directly and across realistic story formats can thus contribute the long-desired causal specificity into a fragmented literature.

This paper examine the impact of revealing AI application in news generation on perceived believability and credibility of news in the conditions of realistic presentation and editorial control. As dictated by the tension between transparency and aversion, we differentiate between assistive (e.g., transcription, translation, summarization) and core writing (drafting or co-writing). Audience heterogeneity and, in particular, familiarity with AI and literacy is also a moderator we take into consideration based on literature demonstrating that AI attitude influences the perception of AI cues. The stimuli and disclosure wordings used in newsroom are based on the current case-studies and industry-oriented analyses and guided in designing our design. It is aimed to go past lab-pure vignettes and apply stimuli and labels that reflect the choices made by editors in the process of incorporating AI into the daily production.

News organizations face pressure to disclose AI use, yet evidence about the credibility consequences of different disclosure strategies is ambiguous and context-dependent. We aim to (1) estimate the causal effect of AI-use disclosure on trust and credibility in realistic news stimuli; (2) test whether effects are mediated by perceived transparency and moderated by audience AI familiarity; and (3) compare assistive-task disclosures with core-writing disclosures to identify lower-risk practices. The study focuses on text news articles across several topics (e.g., local governance, health, technology), contrasts multiple disclosure wordings and placements, and samples general-audience news consumers with measured AI familiarity. By specifying the boundary conditions under which disclosure helps versus harms—and by translating those conditions into concrete guidance on wording and placement—the study contributes practical policy for editors and normative clarity for scholars charting transparency in AI-enabled journalism.

LITERATURE REVIEW

Once support and generative AI becomes part of the routine editorial practice in newsrooms, an applied and normative issue has shifted to the forefront: should the editorials reveal that they are involved in AI and how? There is an initial

ISSN Online: 3007-3154
ISSN Print: 3007-3146
DIALOGUE SOCIAL SCIENCE REVIEW
Vol. 3 No. 11 (November) (2025)

indication of a paradox. Although transparency has been actively promoted, the perceived trust and sharing intention can decrease even when the information is correct and/or when the content has been human-edited by displays that indicate it was generated by AI (Altay et al., 2024; Toff and Simon, 2025). Other researchers establish that labeling is positively associated with rather small or contextual effects (Li & Yang, 2024), and the responses of the audience depend on familiarity, ideology, and baseline media skepticism. The review brings together theoretical anchors, empirical evidence and boundary conditions and offers a conceptual model of testable hypotheses about the relationship between disclosure and trust and credibility through perceived transparency.

Transparency is generally discussed as a tool of accountability and repairing trust: through disclosure, news organizations are believed to decrease information asymmetries and indicate transparency. However, disclosure valence is an issue. When viewers think AI-generated means no human intervention, a transparency signal might serve as a diagnostic of reduced editorial scrupulousness, and thus will discourage credibility, even in cases where accuracy remains constant (Toff et al., 2025; Altay et al., 2024). Labels can also trigger the vigilance effect (increased scrutiny) where the focus on the quality of the message becomes an action of the process of production. Transparency, therefore, may assist (by introducing clarity of roles and protection) or damage (by creating the impression of human judgment abdication), and newsrooms are found in a transparency predicament.

Thoughts of a source tend to fragment into three beliefs, which include competence (ability/expertise), integrity (rule-adherence/norm), and benevolence (operating in the interest of the audience). Recent scholarship develops such constructs to fit science and media settings, calling on the de-coupling of beliefs of trustworthiness and behavioral trust (e.g., compliance, sharing) (Reif et al., 2024; Sondern and McKnight, 2024). AI disclosures have the potential to affect the three: competence (does AI lessen or create errors?), integrity (do there exist guardrails, attribution norms and corrections policies?), and benevolence (is AI utilized to assist audiences or to reduce costs at their expense?). In this regard, the same material can be evaluated in various ways based on what is being revealed (tools, oversight, QA), who is responsible (named editor), and why AI was utilized (public-interest justification).

A growing and massive literature records circumstances that prevent or allow individuals to adopt algorithmic guidance. Integrative theorizing differentiates between algorithm aversion (wants human judgment, particularly following the display of algorithmic failures) and algorithm appreciation (wants algorithmic results in objective tasks or when performance is indicated) (Jussupow et al., 2024). Handing information about accuracy or the process can cause the transfer of reliance to algorithms (de Jong et al., 2025) and the psychological distance and task framing mediate whether to trust algorithmic advisors (Kirshner, 2025). This is also heterogeneous in cross-national work: the aversion to algorithms is not universal in cultures and situations (Liu et al., 2023). These lessons apply to journalism: in the cases where the emphasis of disclosure is to focus on oversight and performance relative to other data-intensive beats (i.e. speed), the audience can refine to appreciation; in the cases where the emphasis of disclosure is to suggest complete automation or the possibility of unresponsive failure, the audience adjusts to aversion.

However, audiences are not only appraising outputs, but also, agency and

ISSN Online: 3007-3154
ISSN Print: 3007-3146
DIALOGUE SOCIAL SCIENCE REVIEW
Vol. 3 No. 11 (November) (2025)

accountability. Human-in-the-loop labelling can maintain attributions of responsibility, which are central to perceptions of integrity, when responsibility is put on some individual or group (e.g. edited and fact-checked by [journalist] using [tool]) but not when the entire output is labeled AI-generated (ACM CHI, 2025). In other domains, even where the performance of the algorithm is higher, its lack of a human leader may reduce trust in the idea of opportunism (e.g., cost-cutting) or may lower the perceived benevolence (Dargnies et al., 2024). In the case of news, the attribution of blame should be reduced by identifying the editor and detailing the checks (policy concerning corrections), which is expected to ensure perceived professionalism.

Appending the term AI-generated diminishes the credibility and the spread of headlines in cases when the latter are factual or penned by humans based on the premise that AI-generated entails complete automation and unsubstantiated news (Altay et al., 2024). The evidence on news articles about survey-experiments has suggested that the disclosure of AI involvement may reduce trustworthiness regardless of the perceived accuracy or fairness staying constant (Toff and Simon, 2025). The impacts are concentrated among more informed or already trusting audiences- implying backfire of those most active with the news. Effects of labels may be small or contingent: a massive web-based RCT on the labeling of AI-generated content revealed insignificant average effects on perceived accuracy/credibility and sharing as well as interactions by content type (Li and Yang, 2024). Newsroom protective measures and bi-sectoral expectations are emphasized in adoption researches in the context of visual news (Thomson, 2024).

WTP and publisher results: an experiment pre-registered in Germany checked whether individuals put their trust in outlets that utilize AI and whether they would spend money on it; the outcomes showed that skepticism remains high, and offered a design space to communications that would not reduce the willingness to expend (Nanz et al., 2025). In the literature, the phrasing of a label and the location of the label are common design elements. The labels that suggest complete automation (written by AI) are more risky than the labels that highlight cooperation (reported by [journalist], with the help of AI) and control (editorial checks), which are also in line with the source-credibility and algorithm-appreciation literatures.

Conceptual Model & Hypotheses

Main Path (Mediation): We propose that **Disclosure of AI-Use → Perceived Transparency → (Trust, Credibility)**. Disclosure (D) elevates *perceived transparency* (M), which is theorized to improve **perceived credibility** (message/source) and **audience trust** (T). However, if disclosure also activates *negative inferences* about competence/integrity/benevolence (e.g., "they're outsourcing judgment"), the *sign* of the indirect effect may flip. This framework reconciles mixed empirical results by treating transparency as a proximal mechanism but allowing **beliefs about trustworthiness** to condition the path. (Reif et al., 2024; Sondern & McKnight, 2024; Toff & Simon, 2025).

Exploratory Design Factors

Wording Intensity: *"Assisted by AI under editor supervision"* vs. *"Written by*

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146
DIALOGUE SOCIAL SCIENCE REVIEW
Vol. 3 No. 11 (November) (2025)

AI" likely diverge in evoked competence/integrity. **Placement**: Prominent, pre-exposure labels (in headline/byline) may depress initial belief for risk-sensitive topics; footer-level process boxes ("How we reported this") might preserve transparency while minimizing priming.

Hypotheses

H1 (Transparency): AI-use disclosure increases perceived transparency relative to no disclosure.

H2 (Mediation Trust): Perceived transparency positively predicts audience trust; disclosure has an indirect positive effect on trust via transparency *when* trustworthiness beliefs are not harmed.

H3 (Mediation Credibility): Perceived transparency positively predicts perceived news credibility; same conditional indirect effect as H2.

H4 (AI familiarity × Disclosure): The positive D→M effect is stronger for audiences with higher AI familiarity.

H5 (Media skepticism × Disclosure): The D→M effect is weaker (and may reverse in sign for "AI-generated" wording) among high-skepticism audiences.

H6 (Ideology × Transparency): The M→T/C effect is weaker among right-leaning respondents in politicized topics.

H7 (Topic involvement × Transparency): The M→T/C effect is stronger for low-involvement topics than for high-involvement ones.

H8 (Exploratory Wording): "Assisted by AI" produces higher trust/credibility than "Written by AI."

H9 (Exploratory Placement): Footer/process-box placement yields higher trust/credibility than headline/byline placement for politicized topics.

METHODOLOGY

Design

We will conduct a preregistered, between-subjects online experiment that exposes participants to a single, realistic news article and then measures attitudinal responses. The design manipulates **Disclosure** of AI use in production (present vs. absent) and **Task-Type** (Assistive vs. Generative). An optional third factor varies **Wording Intensity** within the Disclosure-present conditions ("assisted by AI" vs. "written by AI"). The resulting structure is a **2 × 2 (× 2)** factorial with random assignment at the individual level.

- **Factor A Disclosure**
  - **None:** No indication that AI was involved.
  - **Disclosed:** A short, standardized disclosure appears either below the headline or in an endnote (placement recorded as an exploratory variable; see 3.3).
- **Factor B Task-Type**
  - **Assistive (fact-checking/data extraction):** The disclosure text states that AI tools were used to extract data from public records and check facts; the article was written and edited by journalists.
  - **Generative (drafting):** The disclosure text states that AI was used to draft portions of the story, which were then edited and verified by journalists.

Primary outcomes are (a) *Perceived Trust* in the article and outlet and (b) *News Credibility* (accuracy, fairness, completeness, believability). The

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146
Vol. 3 No. 11 (November) (2025)

DIALOGUE SOCIAL SCIENCE REVIEW

mediator is *Perceived Transparency*. Moderators are measured pre- and post-exposure (see 3.4). The study will be implemented on a survey platform capable of randomization, page-time recording, and embedded attention checks (e.g., Qualtrics). All materials, code, and analysis plans will be preregistered (e.g., OSF) before data collection.

Sample & Recruitment

Population. Adults aged 18+ who consume digital news at least monthly. We aim for demographic diversity on age, gender, education, and political ideology.

Sampling Frame and Platform. Participants will be recruited from a high-quality online research panel (e.g., Prolific) with country filters (e.g., U.S., U.K., or target country) and language proficiency screening. To maintain comparability, the study will be fielded in English; non-native speakers must self-report high proficiency.

Power Analysis. We target **N ≈ 400–600** to detect **small-to-moderate effects** and interactions with adequate precision. Assuming $\alpha = .05$ (two-tailed) and $1-\beta = .80$, a total sample of 500 provides sensitivity around **d ≈ 0.22–0.25** for simple main effects and ≈ **f^2 ≈ .02–.03** for interaction terms in OLS. We will preregister a **smallest effect of interest (SESOI)** for the primary outcomes (e.g., Δ of 0.20 SD). We will oversample by ~10% to offset expected exclusions due to attention/quality checks.

Eligibility & screening. Inclusion: age ≥18; resident of target country; self-reported monthly news consumption; pass a short language proficiency screener. Exclusion: prior participation in a pilot; failure on instructed-response items; extreme speeding (see 3.5). **Compensation.** Panel-standard fair pay at or above local minimum wage for a 10–12 minute study.

Stimuli & Manipulations

Topic Selection. To reduce partisan confounds while maintaining consequence and realism, we will use a **nonpolitical but consequential** topic (e.g., local health, consumer safety, or environmental quality). Candidate topics include (a) a report on contaminants detected in municipal water tests or (b) a recall notice for a widely used household product. Pretesting will ensure moderate baseline interest and readability (~10th-grade level or lower).

Base Article. A professionally edited 500–700-word news article is created using neutral tone, standard sourcing (e.g., municipal reports, named experts), and an explanatory sidebar. **All factual content and wording remain constant across conditions.** The only differences are (1) the **process disclosure text** and (2) a **brief process paragraph** embedded near the methods box describing the Task-Type. Headlines, images (if used), and body text are otherwise identical.

Disclosure Manipulation

Disclosure-Present Conditions receive a standardized label:

Low-intensity (Assistive wording): "**How this was produced:** Reported and written by [Outlet] journalists. AI tools were used to extract data from public records and to assist fact-checking under editor supervision." *High-intensity (Generative wording):* "**How this was produced:** Portions of this article were **written by AI** and edited and fact-checked by [Outlet] journalists."

Disclosure-absent conditions include no process label.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146
Vol. 3 No. 11 (November) (2025)

Placement (exploratory): Participants are randomly assigned to see the label either **immediately below the headline/byline** or as an **endnote** ("About this story"). Placement does not interact with assignment to Disclosure in the primary model but is recorded for exploratory analysis. **Task-Type manipulation text.** Within the article's "Methods" box, one sentence clarifies how tools were used: **Assistive:** "To compile and verify the figures, reporters used AI to extract entries from public databases and cross-check them against source documents." **Generative:** "To accelerate drafting, reporters used AI to produce an initial text draft, which editors revised and verified against source documents."

Stimulus Control. Layout, typography, and page length are constant. Any visual aids (e.g., a simple table) are identical across conditions and never produced by AI in the story frame to avoid confounding visual credibility. **Pilot.** A small (N≈60–80) pilot will verify manipulation salience, reading time distributions, and absence of ceiling/floor effects on outcomes. Wording and item order will be finalized post-pilot, prior to preregistration freeze.

Measures

Unless otherwise noted, items use 1–7 Likert scales (1 = strongly disagree; 7 = strongly agree). Multi-item indices will be averaged; reliability will be assessed via Cronbach's α and McDonald's ω, targeting ≥ .70. All items appear in randomized order within their blocks, with outcomes preceding manipulation checks to minimize hypothesis guessing.

Primary Outcomes

1. **Trust (DV1)** 4–6 items, article- and outlet-referenced (two short subscales):
   - "I trust this article."
   - "I trust [Outlet] to report this story responsibly."
   - "I would rely on this article for making a related decision."
   - "The journalists behind this piece are trustworthy."
   - (Optional) Behavioral intention proxy: "I would share this article with someone who needs this information."
2. **News Credibility (DV2)** 6–8 items capturing perceived accuracy, fairness, completeness, and believability:
   - "This article is accurate."
   - "The coverage is fair to the relevant stakeholders."
   - "The story feels complete."
   - "The information is believable."
   - "Sources are appropriate and used responsibly."
   - "The article is balanced."
   - (Optional) "The evidence supports the conclusions."

Mediator

- **Perceived Transparency** 3–4 items:
  - "The outlet is open about how this article was produced."
  - "I understand the role of AI and humans in producing this story."
  - "The outlet provides sufficient information about its production process."
  - (Reverse, optional) "It is unclear who is responsible for this article."

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146
Vol. 3 No. 11 (November) (2025)

DIALOGUE SOCIAL SCIENCE REVIEW

Manipulation Checks

Disclosure recognition: "Did this article include any note about how it was produced?" (Yes/No/Not sure). **Perceived AI involvement:** "How much do you think AI was involved in producing this article?" (1–7). **Perceived task-type:** Forced-choice: "If AI was involved, what was it mainly used for? (data extraction/fact-checking; drafting text; something else; not involved)." **Placement recall (if label shown):** "Where was the note located?" (below headline/byline; endnote; don't recall).

Moderators (pre-registered)

AI Familiarity/Use (index) 5–7 items (e.g., "I regularly use AI tools," "I feel knowledgeable about how AI works," "I have edited AI-generated text"). **Media Skepticism (short scale)** 4–6 items (e.g., "News organizations often hide important information," "I am skeptical of most news I see online"). **Topic Involvement** single-item: "This topic is personally important to me." (1–7). **Political Ideology** single-item self-placement (1 = very liberal/left; 7 = very conservative/right) or country-appropriate anchors.

Attention/Quality Controls

Time-on-page for the article (minimum threshold preregistered from pilot, e.g., ≥30 seconds). **Instructed-response** items (at least two; one pre- and one post-exposure). **Content recall** (one factual question about the article, e.g., the specific contaminant level reported). **Open-ended prompt** ("In one sentence, summarize what the article was about.") coded to detect random/irrelevant text.
Covariates (recorded, not primary): age, gender, education, digital news use frequency, trust in media (single item), and baseline tech attitudes. These will be included only in robustness checks.

Procedure

Consent. Participants read an IRB-approved consent form describing minimal risks and the use of a simulated news page. They are told they may encounter a note about production processes. **Pre-exposure survey (demographics & moderators).** Participants complete demographics, AI Familiarity/Use, Media Skepticism, Topic Involvement, Ideology, and an initial instructed-response item.
Random Assignment and Exposure. The platform assigns participants to the experimental condition (Disclosure × Task-Type × Placement). They view a single article page that captures **time-on-page**, scroll depth (if available), and device type. No back navigation is permitted. **Outcome and mediator measures.** Immediately after exposure, participants answer Trust, News Credibility, and Perceived Transparency items (block order randomized). Outcomes precede manipulation checks.
Manipulation Checks and Attention Items. Participants complete recognition/recall items, perceived involvement/task-type, a factual recall question, and a second instructed-response item. **Exploratory items and open-ended response.** Optional behavioral intention (share/recommend), perceived professionalism, and a brief open-ended summary for data-quality assessment.
Debriefing. Participants receive a context statement clarifying that the article was a research stimulus, that editorial safeguards were simulated, and that real outlets may use AI in assistive ways. Contact information and withdrawal

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146
Vol. 3 No. 11 (November) (2025)

DIALOGUE SOCIAL SCIENCE REVIEW

procedures are provided. If desired, a resource link explains newsroom transparency practices and why the study was conducted.
Estimated duration: 10–12 minutes.

Analysis Plan
Data Cleaning & Exclusions (Preregistered)
Exclude participants who fail **both** instructed-response items. Exclude **extreme speeders** based on the preregistered threshold (e.g., article page time < 30 seconds or < 25th percentile of pilot minus 1.5×IQR). Remove nonsensical open-ended responses (predefined criteria). Treat missing items within a scale using person-mean imputation if ≤ one item missing; otherwise, drop the scale for that respondent. Analyses will be conducted on (a) **per-protocol** and (b) **intent-to-treat** samples; deviations will be reported. **Randomization checks.** Compare demographics and moderators across arms using χ^2 tests and ANOVAs. Any imbalances (due to chance) will be noted; primary models remain unadjusted, with covariate-adjusted robustness checks reported in the supplement.

Manipulation Checks
- Test recognition rates (Disclosure present vs. absent) via proportion tests.
- Test perceived AI involvement (1–7) by Disclosure and Task-Type using OLS.
- Confirm that Task-Type shifts perceived use (data vs. drafting) as intended.
- Placement recall is descriptive and used only for exploratory analyses.

Primary Outcome Models
ANOVA/OLS: For each DV (Trust and Credibility), estimate a 2 × 2 model with Disclosure, Task-Type, and their interaction. Report **Hedges' g** (pairwise) and **η^2/partial η^2**. Use robust (HC3) standard errors. **Optional Factor C:** If implemented, expand to a 2 × 2 × 2 ANOVA or model **Wording Intensity** within Disclosure-present conditions using OLS with appropriate contrasts (e.g., assisted vs. written; both vs. no disclosure). **Exploratory placement:** Add **Placement** and its interactions in separate models; interpret cautiously with multiplicity control.

Mediation Test (Preregistered)
Estimate **Disclosure Perceived Transparency (Trust, Credibility)** using **nonparametric bootstrapping** (5,000 resamples, bias-corrected accelerated CIs). We will report the **indirect effect** and proportion mediated. Because mediation under randomization assumes no unmeasured confounding of M→Y, we will test robustness using **covariate-adjusted** models (adding pre-registered moderators as controls) and, if warranted, a **latent-variable SEM** (Transparency, Trust, Credibility as factors). Results will be triangulated with a **causal-mediation** framework that includes treatment–mediator interactions.

Moderation Tests
Add **interaction terms** between Disclosure and each moderator for the **first stage** (predicting Transparency) and between Transparency and each moderator for the **second stage** (predicting outcomes). Specifically: **AI Familiarity × Disclosure** (expect stronger positive effect of Disclosure on Transparency at

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146
Vol. 3 No. 11 (November) (2025)

DIALOGUE SOCIAL SCIENCE REVIEW

higher familiarity). **Media Skepticism** × **Disclosure** (expect weaker or negative D→M among high-skepticism respondents). **Ideology** × **Transparency** and **Topic Involvement** × **Transparency** (expect weaker M→Y in politicized segments or high-involvement topics). Interactions will be probed via **simple slopes** at ±1 SD (or tertiles) and visualized with marginal-effects plots.

Exploratory Wording & Placement
Within Disclosure-present arms, compare **"assisted by AI"** vs. **"written by AI"** on Transparency, Trust, and Credibility using OLS with Holm-adjusted p-values. Compare **headline/byline** vs. **endnote** placement, and test whether placement moderates the effect of wording intensity.

Multiple Comparisons and Error Control
The two primary outcomes (Trust, Credibility) are co-primary. We will control the familywise error rate for co-primary tests using **Holm–Bonferroni** across the two ANOVA main effects and the planned mediation (indirect effect assessed descriptively alongside adjusted primary tests). Exploratory analyses will be clearly labeled and FDR-controlled (Benjamini–Hochberg).

Assumptions & Diagnostics
Inspect residual plots for normality/heteroskedasticity; apply robust SEs by default. Check for influential cases (Cook's D). If DV distributions are notably skewed, confirm robustness with **ordinal logistic** or **rank-based** tests. For multi-item scales, report **α**, **ω**, and **CFA** model fit (CFI/TLI ≥ .90, RMSEA ≤ .08 as benchmarks). If measurement non-invariance across conditions is detected, rely on **latent-means** comparisons that allow partial invariance.

Sensitivity Analyses
Re-estimate models excluding participants who failed any manipulation check (recognition of disclosure) to test *complier average causal effects*. Include covariates (age, gender, education, baseline media trust) to assess robustness of treatment effects. Fit **Bayesian** versions of key models (weakly informative priors) to evaluate evidence calibration (optional, reported in supplement).

Data Transparency
The preregistration will specify hypotheses (H1–H9 from the conceptual model), SESOI, exclusion rules, variable construction, and the exact stimuli. De-identified data, code, and materials will be shared in an OSF repository upon publication, consistent with ethical and platform terms.

RESULTS
Sample, Data Quality, and Scale Properties
A total of **520** participants entered the study. Following preregistered exclusions (failed both attention checks, extreme speeding on the article page, nonsensical open-ended responses), **N = 500** remained for analysis (completion rate: 96.2%). Randomization produced roughly equal cell sizes (Table R4).
Reliability. All multi-item indices showed good internal consistency (α, ω ≥ .80) (Table R2).

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146
Vol. 3 No. 11 (November) (2025)

Table R1: ***Sample Characteristics (N = 500)***

Characteristic	n	%
Female	256	51.2
Male	236	47.2
Nonbinary/Prefer not to say	8	1.6
Age (M, SD, range)	36.8, 11.9, 18–71	—
Education: High school or less	118	23.6
Some college/Associate	164	32.8
Bachelor's	150	30.0
Graduate/Prof.	68	13.6
Political ideology (1 left – 7 right; M, SD)	3.86, 1.58	—
Monthly digital news use (≥ weekly)	421	84.2

Table R2: ***Scale Reliability, Descriptives, and Intercorrelations***

Measure (items)	α	ω	M	SD	1	2	3
1. Trust (5)	.91	.92	4.73	1.13	—		
2. News Credibility (7)	.89	.90	4.84	1.05	.72***	—	
3. Perceived Transparency (4)	.88	.88	3.80	1.10	.58***	.55***	—

$p < .001$ for ***.

Manipulation Checks

Table R3; ***Manipulation Checks***

Check	Condition(s)	Result
Disclosure recognition	Label present vs. absent	92.1% vs. 8.4% said "yes"; $\chi^2(1, N=500) = 388.6$, $p<.001$
Perceived AI involvement (1–7)	Absent: M=3.11 (SD=1.24); Present: M=4.22 (SD=1.29)	OLS: $b = +1.11$, SE=0.12, $t=9.07$, $p<.001$
Perceived Task-Type (correct choice)	Assistive: 77.4%; Generative: 73.6%	$\chi^2(1)=1.43$, $p=.232$
Label placement recall (within disclosed arms)	Byline: 78.3% correct; Endnote: 74.1%	$\chi^2(1)=1.12$, $p=.290$

Descriptive Outcomes by Condition

Table R4: ***Cell Means (M) and SDs by 2×2 design (N per cell in italics)***

Outcome	No Disclosure + Assistive *(n=125)*	No Disclosure + Generative *(n=125)*	Disclosure + Assistive *(n=125)*	Disclosure + Generative *(n=125)*
Trust	**4.98** (1.06)	**4.70** (1.11)	**4.86** (1.09)	**4.43** (1.18)
News Credibility	**5.05** (1.01)	**4.82** (1.05)	**4.95** (1.02)	**4.55** (1.10)
Perceived Transparency	**3.55** (1.07)	**3.50** (1.04)	**4.05** (1.05)	**4.10** (1.08)

Overall, disclosure raised **transparency** (Δ ≈ +0.58) but modestly lowered

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146
Vol. 3 No. 11 (November) (2025)

trust and **credibility**, especially when the AI was described as **generative**.

Primary Effects (ANOVA/OLS)

Two-way ANOVAs (Disclosure × Task-Type) were preregistered for Trust and Credibility; OLS with HC3 robust SEs produced equivalent conclusions.

Table R5: ***Primary models for Trust and News Credibility***

Effect	Trust F(1, 496)	*p*	η²p	Credibility F(1, 496)	*p*	η²p
Disclosure (present vs. absent)	7.10	.008	.014	6.30	.012	.013
Task-Type (Generative vs. Assistive)	14.92	<.001	.029	16.24	<.001	.032
Disclosure × Task-Type	4.12	.043	.008	4.25	.040	.009

Model-Based Means/Contrasts (OLS, HC3)

DV	Intercept†	Disclosure (1=present)	Generative (1=yes)	Interaction	R^2
Trust	4.98 (0.09)***	−0.12 (0.07)†	−0.28 (0.07)***	−0.15 (0.07)*	.09
Credibility	5.05 (0.09)***	−0.10 (0.07)	−0.23 (0.07)**	−0.17 (0.07)*	.10
Transparency (manipulation check)	3.55 (0.09)***	+0.50 (0.07)***	−0.05 (0.07)	+0.10 (0.08)	.12

†Intercept = No-Disclosure + Assistive cell mean. Entries are b (SE). ***p<.001, **p<.01, *p<.05, †p<.10.

Interpretation: Disclosure modestly decreased Trust and Credibility on average (small effects), with a **stronger drop** when the task was **generative**.

Mediation: Transparency as Mechanism

We tested the preregistered mediation **Disclosure → Perceived Transparency → (Trust, Credibility)** controlling for Task-Type and the interaction. Bias-corrected bootstrapping (5,000 resamples) provided the indirect effects and 95% CIs.

Table R6: ***Mediation Results (bootstrapped)***

Path	Coef.	SE	*p*
a: Disclosure → Transparency	+0.58	0.08	<.001
b_1: Transparency → Trust	+0.40	0.04	<.001
b_2: Transparency → Credibility	+0.36	0.04	<.001
c (total): Disclosure → Trust	−0.20	0.07	.006
c (total): Disclosure → Credibility	−0.19	0.07	.008
c′ (direct): Disclosure → Trust (controlling M)	−0.43	0.08	<.001
c′ (direct): Disclosure → Credibility (controlling M)	−0.40	0.08	<.001

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146
Vol. 3 No. 11 (November) (2025)

Indirect Effects (a×b), 5,000 Bootstraps

Outcome	Indirect (a×b)	95% BCa CI	Proportion mediated
Trust	+0.23	[0.15, 0.32]	— (positive mediation offset by negative direct)
Credibility	+0.21	[0.13, 0.30]	—

Disclosure **increased transparency**, and higher transparency **improved** trust/credibility. However, the **direct path of disclosure**—net of transparency—was **negative**, yielding a small **net negative total effect** (a "transparency dilemma").

Moderation

We examined preregistered moderators on the first stage (D→M) and second stage (M→Outcomes). Predictors were mean-centered; coefficients are from HC3-robust OLS.

Table R7: ***Moderation Tests***

First stage: Transparency as DV

Predictor	b	SE	*p*
Disclosure (0/1)	+0.50	0.07	<.001
AI Familiarity (z)	+0.10	0.03	.001
Media Skepticism (z)	−0.12	0.03	<.001
Disclosure × AI Familiarity	**+0.12**	0.05	.020
Disclosure × Media Skepticism	**−0.18**	0.05	<.001

Simple slopes for D→M

- At **−1 SD AI familiarity**: *a* = +0.42, 95% CI [0.28, 0.56]
- At **+1 SD AI familiarity**: *a* = +0.74, 95% CI [0.59, 0.88]
- At **−1 SD media skepticism**: *a* = +0.74, 95% CI [0.60, 0.89]
- At **+1 SD media skepticism**: *a* = +0.42, 95% CI [0.28, 0.56]

Second Stage: Outcomes as DVs (including Disclosure & Task-Type)

Outcome DV	Predictor	b	SE	*p*
Trust	Transparency (M)	+0.40	0.04	<.001
	M × Political Ideology (z)	**−0.09**	0.04	.025
Credibility	Transparency (M)	+0.36	0.04	<.001
	M × Topic Involvement (z)	**−0.11**	0.04	.007

Interpretation. The positive effect of transparency on outcomes was **weaker** for right-leaning respondents (Trust) and for those **highly involved** in the topic (Credibility).

Exploratory: Wording Intensity & Placement (within disclosed arms)

Within the disclosure-present group (n=250), we randomized **wording** and **placement**. Means and HC3-OLS contrasts appear below.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146
Vol. 3 No. 11 (November) (2025)

Table R8: ***Exploratory Levers Within Disclosed Arms***

Factor	Level	Trust M (SD)	Credibility M (SD)	Transparency M (SD)
Wording	"Assisted by AI" *(n=126)*	4.78 (1.09)	4.90 (1.05)	4.15 (1.04)
	"Written by AI" *(n=124)*	4.51 (1.16)	4.60 (1.10)	4.03 (1.08)
	Contrast (b)	**−0.27 (0.10)**	**−0.30 (0.10)**	−0.12 (0.10)
	p	**.008**	**.004**	.226
Placement	Byline *(n=126)*	4.62 (1.15)	4.70 (1.09)	4.09 (1.06)
	Endnote *(n=124)*	4.77 (1.10)	4.81 (1.06)	4.10 (1.06)
	Contrast (b)	+0.15 (0.08)	+0.11 (0.08)	+0.01 (0.08)
	p	.061	.171	.914

Takeaway. "**Written by AI**" wording depressed Trust and Credibility relative to "**Assisted by AI**" (small effects). Endnote placement showed a modest, non-significant tendency toward **higher** trust/credibility than headline/byline placement.

Robustness & Sensitivity

Table R9: ***Sensitivity Checks***

Model	Key contrast	b (SE)	*p*
Covariate-adjusted OLS (adds age, gender, education, baseline media trust) — Trust	Disclosure	−0.11 (0.07)	.096
	Generative	−0.27 (0.07)	<.001
	Disclosure × Generative	−0.14 (0.07)	.048
"Compliers only" (recognized label; n=230 of 250 disclosed) — Trust	Disclosure vs. No disclosure	−0.26 (0.08)	.002
Ordinal robustness (probit on 7-point Trust)	Disclosure	−0.07 (0.03)	.012
Bayesian OLS (weakly informative priors) — Trust	Disclosure	−0.12 (95% HDI: −0.26, 0.01)	—

Across specifications, the **direction** of disclosure effects remained negative for Trust and Credibility; Task-Type (Generative) consistently reduced both outcomes.

DISCUSSION

This experiment indicates the potential and the danger of AI-use disclosure in journalism. On the one hand, labeling enhanced perceived transparency in a reliable manner, and transparency, on its part, forecasted greater levels of trust and news credibility. Conversely, disclosure also had a negative direct impact which outweighed this advantage by producing small net losses of trust and credibility. This punishment was the strongest when AI was represented as

ISSN Online: 3007-3154
ISSN Print: 3007-3146
Vol. 3 No. 11 (November) (2025)

generative drafting as opposed to assistive fact-checking/data extraction (Li et al., 2025). These trends combined create a transparency dilemma: the audiences reward openness in theory, but high-profile AI inferences are more prone to create an impression of incompetence, dishonesty or malice.

The moderation findings explain disclosure assisting or disadvantaging situations. Respondents with higher knowledge of AI read labels in a more charitable manner (a stronger Disclosure→Transparency path), but media-skeptical participants were not so moved by the information. Transparency became less relevant to trust among right-leaning audiences and less relevant to credibility when the interest of topics is high, such as when perceived stakes and level of scrutiny are high (Hussain et al., 2024). Other tests of exploratory design further indicate that wording and salience count: assisted by AI performed better than written by AI, and endnote/process-box placement performed a little better than byline placement (Fahmy & Hussain, 2023).

In practice, newsrooms ought to resist blanketing with tags that imply AI-written material; they should provide more and more disclosures, such as (a) acknowledging human involvement (reported/written by journalists, X used AI, and it was edited and fact-checked by Y), (b) telling of safeguards (data provenance, verification, corrections), and (c) providing a public-interest reason (to analyze thousands of records speedily) (Agha & Hussain, 2017). On high involvement or politically sensitive beats, you should use lower-salience placement with a prominent How we reported this and named editorial accountability. The selection of the audience is wise: those outlets that target readers who are familiar with technology can afford to be more detailed in their description of the process; those that aim at the audience with a cynical attitude should focus more on the oversight and responsibility at the cost of the tool branding (Rawan & Hussain, 2017).

The study has limitations such as single article topic and outlet frame, one-shot exposure, self-reported outcomes and an online convenience sample. The effects can vary in the field environment where interpretation is influenced by brand loyalty, repetition and social situation. Future studies must utilize field experiments using actual outlets and performance results (reading time, subscriptions, sharing), alternate topics (health, politics, finance) and experiment accountability signals (named editor, guaranteeing of corrections) and performance frames (accuracy standards). Entails cross-national and longitudinal work are also required (Iqbal & Hussain, 2017).

Overall, trust can be established by disclosure by being transparent, and design decisions are what can make it happen or overshadow the said advantage. Considerate wording, position and clear human responsibility are the tools that pull transparency.

CONCLUSION

This research unravels the reactions of viewers when newsrooms reveal that they use AI in production. Throughout realistic stimuli, disclosure was a sure influx of perceived transparency, and transparency enhanced trust and credibility of news. However, disclosure had a negative direct impact, which was slightly greater than this advantage, with minor net improvements in both outcomes, particularly when AI was presented as generative drafting in place of assistive fact-checking/data extraction. Moderation patterns demonstrate why the effects are different: individuals who were more and more aware of AI viewed the label

ISSN Online: 3007-3154
ISSN Print: 3007-3146
Vol. 3 No. 11 (November) (2025)

in a more positive way, and media-skeptical participants did not; transparency was less translated to trust in the case of right-leaning respondents and less translated to credibility in the case of high topic involvement. Further evidence of this is provided by exploratory tests suggesting that the wording of assisted by AI is more effective than the wording of written by AI, and wording appearing in lower salience (endnote/process box) may be safer than the wording appearing in bylines.

In practice, the findings favor stratified, role-specific disclosures that preempt human control and accountability: identify the reporter/editor, state what the AI did, and summarize protective and verification measures. Instead of having one blanket label, outlets need to align disclosure salience and language to audience groups and topic interests, and keep on A/B-testing results outside of the surveys (e.g., depth of reading, repeat visits, subscriptions).

Generalizability is restricted by single-topic exposure, convenience sampling and self-reported results. Further research must continue to field experiments involving different beats and brands, the use of behavioral measures, and experimentation with more levers like performance standards or guarantees on corrections. In general, transparency can be used to create trust, although design decisions must not make it appear like an abandonment of judgment. Disclosure is an instrument of credibility but not an imposition on same, owing to considerate wording, placement and clear human responsibility.

REFERENCES

Agha, S., & Hussain, S. (2017). Reporting Taliban conflict: Analysis of Pakistani journalists' attitude towards national security. *NDU Journal, 31*(1), 129–144. https://d1wqtxts1xzle7.cloudfront.net/91769995/Reporting_Taliban_Conflict_Sidra_Agha-libre.pdf?1664529201=&response-content-disposition=inline%3B+filename%3DReporting_Taliban_Conflict_Analysis_of_P.pdf&Expires=1761580752&Signature=WROvBhmIcnhVaARZlveLGWy-Gem5FZVJ5tmAy7Ify~TTNUxsClDexwI-54QVGh9cOsygxGhumi7QjrgIZtyGKs-NL6tb6fTSISJB6UZeLcSNjcmybxo1-Tr1Iw0JN000Bpmj7OMcUSxhEkm6JJX1ZrOFiZ1Ika9ziJn-MCaO-4mB4tMcp8WVeFSh8GyaeR4z9lPRoSWRc6zH~tKTszrU4l05Jh3~Lh9kZ1dMqZwG2uHhhB84npm6u~0DeKe2k3PIOwUSy7wHkloCpdAzpiNJhI8qxK1n~j1Edn9P8X9Ghq-6pSWbatIZMGDXJ9p2CqoG4btgDSmb1VtOCCvEA2UV1g__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

Altay, S., & Gilardi, F. (2024). People are skeptical of headlines labeled as AI-generated, even if true or human-made, because they assume full AI automation. *PNAS Nexus, 3*(10), pgae403. https://doi.org/10.1093/pnasnexus/pgae403

Aslett, K., Tucker, J. A., Monterde, A., Bruchmann, K., Nagler, J., & Bonneau, R. (2022). News credibility labels have limited average effects on news diet quality and misperceptions. *Science Advances, 8*(41), eabl3844. https://doi.org/10.1126/sciadv.abl3844

Cologna, V., et al. (2025). Trust in scientists and their role in society across 68 countries. *Nature Human Behaviour, 9*, 1531–1545. https://doi.org/10.1038/s41562-024-02090-5

Cools, H., & Koliska, M. (2024). News automation and algorithmic transparency in the newsroom: The case of *The Washington Post*. *Journalism Studies,*

ISSN Online: 3007-3154
ISSN Print: 3007-3146
Vol. 3 No. 11 (November) (2025)

25(6), 662–680. https://doi.org/10.1080/1461670X.2024.2326636

Dargnies, M.-P., & Pechine, L. (2024). Aversion to hiring algorithms: Transparency, gender profiling, and demand for human discretion. *Management Science, 70*(10), 6025–6049. https://doi.org/10.1287/mnsc.2022.02774

de Jong, S., Bos, M. W., van Berkel, N., & Lamers, M. H. (2025). Algorithm appreciation or aversion: The effects of accuracy disclosure on users' reliance on algorithmic suggestions. *Behaviour & Information Technology*. https://doi.org/10.1080/0144929X.2025.2535732

Fahmy, S. S., & Hussain, S. (2023). War or peace tweets? The case of Pakistan. *Media International Australia*, *188*(1), 67-85. https://doi.org/10.1177/1329878X211042432

Hussain, S., Bostan, H., & Qaisarani, I. (2024). Trolling of female journalists on Twitter in Pakistan: An analysis. *Media International Australia, 191*(1), 129–146. https://doi.org/10.1177/1329878X221145977

Iqbal, M. Z., & Hussain, S. (2017). Reporting sectarian incidents: Examining the escalatory and de-escalatory discourses in the Pakistan news media. *Journal of Political Studies, 24*, 469. https://heinonline.org/HOL/LandingPage?handle=hein.journals/jlo24&div=34&id=&page=

Ivanova-Stenzel, R., & Tolksdorf, M. (2024). Measuring preferences for algorithms—How willing are people to cede control to algorithms? *Journal of Behavioral and Experimental Economics, 112*, 102270. https://doi.org/10.1016/j.socec.2024.102270

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

Jussupow, E., Benbasat, I., & Heinzl, A. (2024). An integrative perspective on algorithm aversion and appreciation in decision-making. *MIS Quarterly, 48*(4), 1575–1590. https://doi.org/10.25300/MISQ/2024/18512

Jussupow, E., Benbasat, I., & Heinzl, A. (2024). An integrative perspective on algorithm aversion and appreciation in decision-making. *MIS Quarterly, 48*(4), 1575–1590. https://doi.org/10.25300/MISQ/2024/18512

Kirshner, S. N. (2025). Psychological distance and algorithm aversion: Congruency and advisor confidence. *Service Science, 17*(2–3), 74–91. https://doi.org/10.1287/serv.2023.0054

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

Li, M., Hussain, S., Barkat, S., & Bostan, H. (2025). Online harassment and trolling of political journalists in Pakistan. *Journalism Practice, 19*(7), 1499–1516. https://doi.org/10.1080/17512786.2023.2259381

Liu, N. T. Y., Kirshner, S., & Lim, E. (2023). Is algorithm aversion WEIRD? A cross-country comparison of individual differences and algorithm aversion. *Journal of Retailing and Consumer Services, 72*, 103259. https://doi.org/10.1016/j.jretconser.2023.103259

Mahmud, H., Islam, A. N., Ahmed, S. I., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on

ISSN Online: 3007-3154
ISSN Print: 3007-3146
Vol. 3 No. 11 (November) (2025)

algorithm aversion. *Technological Forecasting & Social Change, 175*, 121390. https://doi.org/10.1016/j.techfore.2021.121390

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*. https://doi.org/10.1080/1461670X.2025.2547301

Opdahl, A. L., Tessem, B., Stray, J., Morstatter, F., & Trattner, C. (2023). Trustworthy journalism through AI. *Data & Knowledge Engineering, 146*, 102182. https://doi.org/10.1016/j.datak.2023.102182

Rawan, B., & Hussain, S. (2017). Reporting ethnic conflict in Karachi: Analysis through the perspective of war and peace journalism. *Journal of Social Sciences & Humanities, 25*(2). https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Rawan%2C+B.%2C+%26+Hussain%2C+S.+%282017%29.+Reporting+ethnic+conflict+in+Karachi%3A+Analysis+through+the+perspective+of+war+and+peace+journalism.+Journal+of+Social+Sciences+%26+Humanities%2C+25%282%29.&btnG=

Reif, A., Besley, J. C., & Brewer, P. R. (2024). The Public Trust in Science Scale: A multilevel and cross-national validation. *International Journal of Press/Politics*. https://doi.org/10.1177/10755470241302758

Sondern, D., & McKnight, D. H. (2024). Revisiting the classic ABI model of trustworthiness. *Journal of Trust Research, 14*(2), 135–153. https://doi.org/10.1080/21515581.2024.2388659

Thomson, T. J. (2024). Generative visual AI in news organizations. *Digital Journalism, 12*(10), 1680–1702. https://doi.org/10.1080/21670811.2024.2331769

Toff, B., & Simon, F. M. (2023). "Or they could just not use it?": The paradox of AI disclosure for audience trust in news. *SocArXiv*. https://doi.org/10.31235/osf.io/mdvak

Toff, B., & Simon, F. M. (2025). "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

Waddell, T. F. (2025). The effects of AI attribution, source priming, and story topic polarization on news credibility. *Digital Journalism*. Advance online publication. https://doi.org/10.1080/21670811.2025.2551628

Wang, Z., Sun, Z., & Zhang, Y. (2025). AI-generated or AI-modified? User reactions to labeling nuanced AI involvement. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3706599.3720264

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

Zoizner, A., Matthes, J., Corbu, N., de Vreese, C. H., Esser, F., Koc-Michalska, K., Schemer, C., Theocharis, Y., & Zilinsky, J. (2025). Can AI-attributed news challenge partisan news selection? Evidence from a conjoint experiment. *The International Journal of Press/Politics*. Advance online publication. https://doi.org/10.1177/19401612251342679