

AI-Generated News and Its Impact on Public Trust in Journalism

Sher Baz Khan

Research Fellow, Erich Brost Institute for International Journalism, Technical University Dortmund, Germany. Email: Sherbazkhan2014@gmail.com

Muntaha Sami

M.Phil, School of Media and Communication, Studies, Minhaj University Lahore, Punjab, Pakistan. Email: muntahasami00@gmail.com

Zarwah Nabil*

Masters in Development Journalism, Institute of Communication Studies, University of the Punjab. Corresponding Author Email: nabilzarwah4@gmail.com

ABSTRACT

Generative AI is reshaping how news is produced and encountered, raising fundamental questions about public trust. This study develops a hybrid credibility framework that distinguishes a process-legitimacy path (disclosures about how stories are made) from a provenance-credence path (verifiable signals of media origins and edits). We report results from a preregistered $2 \times 2 \times 2$ between-subjects experiment (Authorship: AI vs. human; Label: present vs. absent; Platform: publisher site vs. social feed; $N=640$) with story type blocked. AI authorship produced a modest trust penalty, larger in feeds and for hard news. Labels substantially increased perceived transparency and slightly improved perceived accuracy, but also carried a small direct negative effect on trust, consistent with algorithm-aversion heuristics; the net effect was positive among high-AI-literacy participants and negative among low-literacy participants. Parallel-mediation analyses showed that perceived accuracy and perceived bias transmitted authorship effects to trust, while transparency mediated label effects. We provide design guidance for contextual disclosure wording (“AI-assisted and editor-reviewed”), layered provenance that travels with assets, and visible editorial oversight across devices and distribution channels online. We conclude that calibrated trust in AI-augmented journalism depends on pairing usable transparency with verifiable provenance and audience literacy supports.

Keywords: AI-Generated News; Trust In Journalism; Source Credibility; Transparency; Provenance; Ai Literacy; Social Platforms.

1. INTRODUCTION

The rapid diffusion of generative artificial intelligence (GenAI) into both newsrooms and the platforms where audiences encounter news has transformed how journalism is produced, distributed, and evaluated. In news organizations, editors and visual desks are experimenting with GenAI to ideate, draft, summarize, translate, illustrate, and localize content at scale; on platforms, GenAI tools increasingly shape what users see and how they judge it, from synthetic images embedded in news feeds to headline rewrites in search. These developments are neither hypothetical nor marginal: comparative studies of photo editors across seven countries document growing, if uneven, adoption of generative visual tools in editorial workflows (Thomson et al., 2024), and recent industry surveys show that the public is aware of, and ambivalent about, AI's role in news production and distribution (Fletcher & Nielsen, 2025). Against this backdrop, questions of credibility, transparency, and professional identity have re-emerged with new urgency (Lewis, 2025).

At the center of these debates is public trust in journalism. While trust in news has been drifting downward in many countries for years (B. Rawan. & S. Hussain., 2017), the advent of GenAI raises the possibility of additional, AI-specific trust dynamics. Audiences may react differently when they believe AI helped write, edit, or illustrate a story; when they see AI-disclosure labels or “content credentials”; or when a platform flags headlines as “AI-generated.” Experimental evidence is mixed and sometimes contradictory. Several studies find that labeling content as AI-generated reduces perceived credibility or willingness to share—even when the labeled content is accurate or human-written (Altay et al., 2024; Toff et al., 2024; Wang et al., 2024). Others report minimal or context-dependent effects (Li et al., 2024), or show that what matters most is not the presence of a label per se but audiences' inferences about the extent of AI involvement (Jia et al., 2024). This pattern mirrors a broader literature outside journalism in which disclosures about AI can backfire, eroding legitimacy and trust (Schilke & Reimann, 2025), yet user reliance on AI advice can remain high in specific decision contexts (Klingbeil et al., 2024). In short, we lack clarity about when, how, and for whom AI cues change trust in news.

This paper addresses that gap by studying **AI-generated news and its impact on public trust in journalism**. We focus on two forms of “AI presence” that audiences actually encounter: (1) *process cues* (e.g., bylines or disclosure statements indicating AI involvement) and (2) *provenance cues* (e.g., verifiable content credentials attached to text,

images, or composites). We ask three interlocking questions: First, what are the **causal effects** of AI process cues on perceived credibility, fairness, and willingness to engage with news? Second, do **provenance cues**—such as authentication signals that indicate how a piece of media was created or altered—mitigate, exacerbate, or leave unchanged the trust effects associated with process cues? Third, how do **prior attitudes and AI literacy** condition these effects? Each question is motivated by emerging evidence but unresolved theory. For instance, studies show that provenance information can reduce trust in deceptive media but also in honest media if the signal is incomplete or invalid (Feng et al., 2023), while usability research suggests that where and how such signals appear on a page strongly shapes their interpretation (Usable News Authentication, 2024).

GenAI's institutionalization in journalism is not merely a technical upgrade; it reconfigures roles, routines, and norms. Visual desks adopt text-to-image tools to illustrate complex stories under deadline; reporters use LLMs to brainstorm angles or draft briefs; editors experiment with AI-assisted copyedits and SEO-oriented headline generation. These practices challenge the profession's longstanding boundary work around authorship, accountability, and transparency (Thomson et al., 2024; Lewis, 2025). Meanwhile, on the audience side, labeling policies proliferate across platforms and publishers. When users encounter an "AI-generated" label on a headline or byline, many infer high AI involvement and lower human oversight, even if production was actually human-led with modest AI assistance (Altay et al., 2024; Toff et al., 2024). The net result is a dynamic environment where **disclosures, inferences, and interfaces** interact to shape trust judgments—often in ways newsroom policies did not intend.

Despite a rush of experimental and field studies, the **causal and associational effects** of

AI involvement on trust in journalism remain unsettled. On one hand, meta-analytic evidence across communication experiments suggests a small but negative average effect of machine authorship on credibility assessments (Wang et al., 2024). On the other hand, domain-specific experiments indicate that credibility penalties can vanish when texts are accurate, well-written, or when AI support is framed as *assistance* to human journalists rather than *authorship* (Henestrosa et al., 2022; Jia et al., 2024). Moreover, the broader human-AI literature has moved beyond generic "trust" to stress **calibrated reliance** the alignment between trust and system quality showing both algorithm

aversion and appreciation depending on context, accuracy, and user priors (Klingbeil et al., 2024). What is missing for journalism is an integrative account that connects *which AI cues are shown*, *where they are shown*, and *what audiences assume* about production, oversight, and verifiability.

Aim and Contributions

Aim. We aim to identify and explain the conditions under which AI-related cues in news **increase, decrease, or leave unchanged** audience trust and to determine how provenance authentication interacts with disclosure labels.

Theoretical Contributions. First, we synthesize transparency-as-legitimacy theory with journalism’s sociotechnical role conceptions to propose a **two-path model of trust**: (a) a *process-legitimacy path* (what audiences believe about how the news was produced) and (b) a *provenance-credence path* (what audiences can verify about media origins and edits). We theorize that AI disclosures often depress legitimacy when they trigger inferences of low human oversight (Schilke & Reimann, 2025; Altay et al., 2024), but that robust, well-placed provenance signals can counteract deceptive media and, in some designs, stabilize trust albeit with risks of over- or under-correction if signals are confusing (Feng et al., 2023; Usable News Authentication, 2024). Second, we extend **role theory in AI-augmented newsrooms**, differentiating authorial, assistive, and curatorial AI roles and predicting distinct trust effects for each (Lewis, 2025; Thomson et al., 2024). Third, we bring **calibrated reliance** into news trust research by specifying when labels should *increase* trust (e.g., when they signal extra scrutiny) and when they should *decrease* it (e.g., when they imply less human accountability).

Methodological Contributions. We outline a multi-method design: (1) preregistered online experiments that manipulate AI process cues (e.g., bylines, end-notes) and provenance cues (e.g., content credentials badges with varying validity) across text and image news formats; (2) interface experiments that vary the *placement and salience* of these cues to estimate UI-mediated effects on trust; and (3) a cross-sectional survey component measuring AI familiarity, literacy, and baseline news trust. Our design explicitly distinguishes **beliefs about AI involvement** from **verifiable provenance**, overcoming a common confound in prior work where “AI-generated” labels were interpreted as claims about content quality rather than production method (Jia et al., 2024; Li et al., 2024).

Practical Contributions. For news organizations, we provide evidence-based guidance on *when to disclose*, *how to phrase and place disclosures*, and *when to prefer explainers over labels*. For platforms and standards bodies, we contribute UI principles for provenance indicators (e.g., iconography, location, and hierarchies of detail) that minimize misinterpretation and improve calibration supporting efforts to authenticate media without unduly penalizing legitimate, AI-assisted journalism (Feng et al., 2023; Usable News Authentication, 2024). For educators and civil society, we specify how **AI and data literacy** shape cue interpretation, informing interventions targeted at groups most susceptible to miscalibration.

Brief Outline of the Paper

The remainder of the article proceeds as follows. **Section 2** reviews literature on trust in news, algorithm aversion and appreciation, and emerging evidence on AI disclosures and provenance signals in journalism. We develop hypotheses linking process and provenance cues to perceived credibility, fairness, and engagement. **Section 3** details the multi-method research design, operationalizing trust outcomes and cue manipulations across text and image formats, and outlines sampling, preregistration, and analysis plans. **Section 4** reports experimental and survey findings, including heterogeneous effects by AI literacy and baseline news trust. **Section 5** discusses implications for theory (transparency-as-legitimacy and calibrated reliance), newsroom policy (disclosure and explainability), and platform design (usable provenance). **Section 6** concludes with limitations and directions for future research, including cross-cultural generalizability and field experiments within publisher CMSs.

Finally, we emphasize a pragmatic stance: AI's presence in journalism will persist and diversify. Rather than presuming that more or less transparency is *always* better, we argue for **contextual transparency** and **usable provenance** that make the *right* information legible at the *right* time, helping audiences calibrate trust to content quality and process integrity (Thomson et al., 2024; Schilke & Reimann, 2025; Feng et al., 2023).

2. Literature Review

2.1 Public Trust in Journalism

Trust in journalism is widely understood as a **multidimensional** set of beliefs regarding (a) the *competence* or ability of news organizations to gather and verify facts; (b) *benevolence*,

i.e., concern for the public interest; (c) *integrity*, i.e., adherence to professional norms; and, increasingly, (d) *transparency* about processes and sources (S. Agha & S. Hussain., 2017; Holtrup et al., 2024; Fletcher & Nielsen, 2025). These dimensions jointly shape whether audiences view particular outlets—and “news” as a category as reliable enough to inform civic decisions. Contemporary survey and panel evidence further suggests that **perceived accuracy and trust are reciprocally reinforcing**: individuals who rate the news they consume as accurate tend to report higher institutional trust in news media, and higher trust can, in turn, color accuracy judgments (Holtrup et al., 2024). At the same time, recent cross-national work documents **contextual asymmetries across platforms**: people generally report lower trust in news encountered on social media feeds than in “news in general,” underscoring that *where* AI-mediated content is encountered matters for downstream credibility assessments (Iqbal, M. Z., & Hussain, S., 2017; Holtrup et al., 2024; Fletcher & Nielsen, 2025).

A second strand highlights the **role of audience priors** ideology, media repertoires, and perceptions of bias—in shaping trust. As generative AI enters the information ecosystem, these priors become even more salient because people may apply *machine* or *automation* heuristics to resolve uncertainty (“If it’s AI, is it biased or sloppy?”), with heterogeneous effects across groups (Toff & Simon, 2025). Emerging work on the news/AI nexus therefore argues that trust should be theorized not just at the *organization* level but also at the *content* and *process* levels (e.g., how stories are made) precisely the layers most directly impacted by AI (Lewis et al., 2025). (Toff & Simon, 2025; Lewis et al., 2025).

2.2 AI in News Production

Newsrooms now deploy AI along a **continuum of automation**:

- **Automated templates** that turn structured data (e.g., sports, finance, weather) into standard write-ups;
- **LLM-assisted drafting** that aids research, outline generation, fact-finding, or copy-editing;
- **Fully synthetic pieces**, where text or visuals are generated end-to-end and edited by humans; and
- **Optimization layers** such as **headline rewriting**, SEO suggestions, and A/B

testing powered by models (Opdahl et al., 2023; Thomson et al., 2024; Lewis et al., 2025). (Opdahl et al., 2023; Thomson et al., 2024; Lewis et al., 2025).

Qualitative studies with editors and visual desks across several countries show that **generative visual AI** is already prompting new policies on disclosure, synthetic image bans outside clearly labeled contexts, and updated verification workflows, while simultaneously creating efficiencies in ideation and illustration (Thomson et al., 2024). Vision papers propose broader AI-enabled pipelines—from source discovery and lead detection to personalization and verification—while insisting on human oversight to maintain editorial accountability (Opdahl et al., 2023). Conceptual work frames these changes as a **structural challenge** to journalism’s institutional coherence (norms, roles, technologies), placing transparency and explainability at center stage for sustaining public trust (Lewis et al., 2025). (Thomson et al., 2024; Opdahl et al., 2023; Lewis et al., 2025).

2.3 Credibility & Transparency

Source Credibility & “Machine Authorship”

Experimental and meta-analytic evidence indicates that **machine/bot/AI bylines can reduce perceived credibility** of news content, although the average effect is small and moderated by topic and context (Wang et al., 2024). In some experiments, explicit AI bylines decreased both message and source credibility (Lermann Henestroso et al., 2024), whereas other studies report minimal or bounded effects depending on how authorship is framed and what audiences infer about human oversight (Jia et al., 2024). Overall, the literature implies that “AI-generated” functions as a **negatively valenced cue** for many users unless offset by cues to editorial control or expertise. (Wang et al., 2024; Lermann Henestroso et al., 2024; Jia et al., 2024).

Labeling and Disclosure

What happens when AI content is **labeled**? Large preregistered experiments find that “**AI-generated**” labels reliably lower perceived accuracy and sharing intent, even for *true* or *human-written* headlines apparently because many readers equate the label with **full automation** and low human oversight (Altay & Gilardi, 2024). Other recent work testing different **labeling strategies** (e.g., process-based vs output-based labels) shows mixed or limited effects on detection and evaluation, underscoring that **label design and placement** matter and that disclosures can have unintended consequences (Wittenberg et

al., 2025). (Altay & Gilardi, 2024; Wittenberg et al., 2025).

Provenance Signals & Content Credentials

Beyond labels, developers and standards bodies have advanced **content provenance** (e.g., embedded edit histories and cryptographic signatures). A multi-country experiment shows that **provenance cues can lower trust in deceptive media** (useful), but also **misfire** when signals are incomplete/invalid, prompting unwarranted skepticism of *authentic* media; user understanding of provenance remains fragile (Feng et al., 2023). This implies that provenance needs careful UI design, literacy scaffolding, and newsroom explainers to avoid “blanket skepticism.” (Feng et al., 2023).

Corrections and Transparency Policies

Transparency also includes **corrections policies** and post-publication accountability. A recent political-communication experiment suggests that **exposure to corrections can increase belief accuracy but reduce trust in the corrected source** highlighting tradeoffs between truth and trust (Freitag et al., 2024). For AI-mediated news, this tradeoff may be sharper if audiences attribute the error to automation rather than to human judgment, strengthening the case for explicit human accountability in disclosures. (Freitag et al., 2024).

Human–AI Collaboration Cues

Finally, transparency about **human-in-the-loop** processes tends to support **appropriate reliance** on AI systems; interactive transparency and co-sourcing (AI + editor) can sustain user trust relative to AI-only conditions in adjacent moderation contexts, suggesting analogous designs for news workflows and story pages (Molina & Sundar, 2022). (Molina & Sundar, 2022).

2.4 Audience Factors

AI Literacy & Media Literacy

How audiences interpret AI cues depends on **AI literacy** knowledge and skills to understand capabilities and limits of AI tools. A 2024 systematic review maps the quality of AI-literacy instruments and emphasizes validated measures such as the **SNAIL** scale; related validation work shows SNAIL captures *technical understanding*, *critical appraisal*, and *practical application* factors (Lintner et al., 2024; Laupichler et al., 2023). These constructs are theoretically adjacent to news literacy and likely moderate how people

react to AI bylines, labels, and provenance. (Lintner et al., 2024; Laupichler et al., 2023).

Political Ideology/Partisanship

AI cues are **not interpreted in a vacuum**: trust responses vary with ideology and prior attitudes toward news and platforms. Experimental disclosure studies document **heterogeneous effects** by partisanship and baseline trust, including defensive processing when automation threatens group priors (Toff & Simon, 2025). This suggests that AI transparency may reduce perceived credibility more among already-skeptical groups, potentially widening *trust gaps* unless coupled with brand-level assurances. (Toff & Simon, 2025).

Platform Use and Encounter Context

Where audiences *encounter* news publisher sites, aggregators, search, or social feeds shapes perceived trust. Recent cross-country analyses show **lower trust on social platforms** compared with direct news use, aligning with findings that platform environments condition accuracy heuristics (Fletcher & Nielsen, 2025). As more AI-composed headlines and summaries circulate via feeds and messaging apps, the **platform context** may magnify or mute the signal of AI labeling and source cues. (Fletcher & Nielsen, 2025).

2.5 Research Gap

Across these strands, three gaps emerge. **First, evidence on labeling is inconsistent** across formats (text vs. visuals), labels (AI-generated vs. AI-assisted/process), and tasks (persuasion vs. accuracy vs. sharing). Some studies find robust negative effects of “AI-generated” labels (Altay & Gilardi, 2024); others report limited or design-dependent effects (Wittenberg et al., 2025), and meta-analytic estimates suggest only **small average penalties** to credibility for machine authorship (Wang et al., 2024). What remains under-tested are **comparative, cross-format** experiments that hold content constant while varying disclosure *type* and *granularity* (e.g., “AI-assisted with human editing” vs. “AI-generated”) and that measure **multiple outcomes** (accuracy, credibility, trust in outlet, willingness to subscribe, etc.). (Altay & Gilardi, 2024; Wittenberg et al., 2025; Wang et al., 2024).

Second, Cross-Platform Generalizability is limited. Most tests occur in single-platform labs or web experiments; few compare the same disclosures across **publisher sites, search, and social feeds**, despite evidence that platform context systematically shapes trust

(Fletcher & Nielsen, 2025). Incorporating **provenance UI** (e.g., content credentials) alongside labels also remains rare, even though provenance can both help and backfire depending on interface and completeness (Feng et al., 2023). (Fletcher & Nielsen, 2025; Feng et al., 2023).

Third, we lack field tests inside newsrooms that pair **workflow transparency** (what AI did, what editors did) with **corrections policy transparency** when AI-assisted stories are updated, to evaluate *longer-term* effects on brand trust. Existing correction experiments show potential **accuracy–trust tradeoffs** (Freitag et al., 2024), but have not been integrated with AI-specific disclosures in ecologically valid settings. Addressing these gaps calls for **pre-registered, cross-format, multi-platform experiments** that manipulate disclosure designs (byline, label, provenance, human oversight) and track multiple audience outcomes over time and across audience segments defined by AI and media literacy. (Freitag et al., 2024).

3. Theoretical Framework & Hypotheses

This study integrates **Source Credibility Theory (SCT)**, **dual-process models of persuasion** (Heuristic–Systematic/Elaboration Likelihood), and **transparency/provenance signaling** to explain when, how, and for whom AI cues in news shape **trust in journalism**. In our framework, two kinds of cues are focal: (a) **process cues** indicating how a story was produced (e.g., “AI-assisted,” “AI-generated,” editor oversight), and (b) **provenance cues** that authenticate origins/edits (e.g., content credentials). We argue these cues act on trust primarily **indirectly** by shifting perceived **transparency, accuracy, and bias** and that effects are **conditional** on audience **AI literacy** and the **encounter context** (platform; story type).

SCT holds that audience judgments of a source hinge on perceived **expertise** and **trustworthiness**, with goodwill/benevolence often added in contemporary media research. Process cues that emphasize human oversight (editor review) are likely to bolster perceived **trustworthiness** and benevolence, while cues implying end-to-end machine authorship may depress perceived **expertise** or accountability (Wang et al., 2024; Altay & Gilardi, 2024). In field-adjacent experiments, AI authorship labels reduce credibility on average, although estimates are small and context-sensitive, consistent with SCT’s prediction that credibility is malleable and multidimensional (Wang et al., 2024). Framing

also matters: “AI-assisted with human editing” plausibly signals retained human **responsibility**, improving integrity attributions relative to “AI-generated” (Jia et al., 2024).

Dual-process accounts clarify **how** such cues work. Under **ELM/HSM**, individuals use either effortful, argument-based evaluation (**systematic/central**) or quicker **heuristic/peripheral** cues (e.g., “AI label $\Rightarrow$ less credible”). Online news is often consumed under low attention, inviting heuristic use (Fletcher & Nielsen, 2025). In that state, a stark “AI-generated” badge can trigger **algorithm-aversion heuristics**, lowering credibility regardless of content quality (Altay & Gilardi, 2024). When issue involvement or ability is higher e.g., among **AI-literate** users who understand what an LLM can and cannot do labels may be integrated **systematically** with other information (Molina & Sundar, 2022; Lintner et al., 2024). Hence, we expect **heterogeneous** label effects by AI literacy: more knowledgeable users should discount crude heuristics and attend to finer distinctions (AI-assisted vs. AI-generated; editor-reviewed vs. not), aligning with “calibrated reliance” perspectives (Wang et al., 2024).

Transparency/provenance signaling adds a **market-for-information** logic: signals that are **costly to fake** (e.g., cryptographically bound content credentials with valid edit histories) can reduce information asymmetries and raise confidence, but **ambiguous or partial signals** can misfire, generating blanket skepticism even toward authentic media (Feng et al., 2023). Usability matters as well: where and how signals are surfaced (badge iconography; hover cards; inline explainer) shape interpretation (Wittenberg et al., 2025). We therefore distinguish **transparency** (audience sense that production is open/legible) from **provenance validity** (verifiable metadata integrity). Process labels should primarily raise **perceived transparency**; provenance should primarily raise **perceived accuracy** but either can backfire if they cue low human oversight or are poorly designed (Altay & Gilardi, 2024; Feng et al., 2023).

Contextual Moderators follow. First, **platform**: trust is systematically lower for news encountered on **social feeds** than on publisher sites, and attention is more fragmented conditions that amplify **heuristic processing** (Fletcher & Nielsen, 2025). We therefore expect stronger negative reactions to crude “AI-generated” labels in feeds, with greater potential value for **provenance cues** that travel with the asset across platforms

(Wittenberg et al., 2025). Second, **story type: hard news** (public-affairs reporting) sets higher stakes for factual accuracy and accountability; audiences may penalize machine authorship more here than for **features** or service journalism where storytelling/aesthetics loom larger (Thomson et al., 2024; Lewis et al., 2025).

Bringing these pieces together, our **conceptual model** posits that process and provenance cues influence **trust in journalism** largely **indirectly** via **perceived transparency**, **perceived accuracy**, and **perceived bias** (lower bias, higher accuracy $\Rightarrow$ higher trust). **AI literacy** moderates cue interpretation, shifting audiences from heuristic discounting to more calibrated evaluations (Molina & Sundar, 2022; Lintner et al., 2024). **Platform** and **story type** shape the processing route and the weight of each mediator.

Research Questions and Hypotheses

RQ1. *How does exposure to AI-generated news (vs. human-authored) affect trust in journalism?*

Grounded in SCT and dual-process expectations, we anticipate average **credibility penalties** for AI-authored content especially in low-attention contexts tempered or reversed when cues make human oversight salient (Wang et al., 2024; Altay & Gilardi, 2024; Jia et al., 2024).

H1 (Mediated Transparency Effect). *Transparent AI labels increase perceived transparency, which in turn increases trust.*

Rationale: Process disclosures satisfy a transparency norm that signals honesty about production, which audiences often reward provided the label does not simultaneously imply **absence** of editorial control. We therefore specify **indirect effects** from label $\rightarrow$ transparency $\rightarrow$ trust, alongside a possible **direct negative** path (label $\rightarrow$ trust) when labels connote full automation (Altay & Gilardi, 2024; Wittenberg et al., 2025).

H2 (AI-Literacy Moderation). *AI literacy moderates label effects; positive transparency-to-trust mediation is stronger among high-literacy users.*

Rationale: High-literacy users are more capable of **systematic** processing, using labels as informative process signals rather than as crude heuristics; low-literacy users are more likely to treat “AI-generated” as a negative **peripheral** cue (Molina & Sundar, 2022; Lintner et al., 2024).

H3 (Accuracy/bias mediation). *Perceived accuracy and perceived bias mediate the relationship between AI exposure and trust.*

Rationale: Under SCT, competence (accuracy) and integrity (lack of bias) are proximal determinants of trust. If AI cues reduce assumed **accuracy** or raise perceived **bias**, trust falls; provenance cues that **verify origins/edits** should raise accuracy perceptions and can offset bias inferences when well designed (Feng et al., 2023; Wang et al., 2024).

RQ2. *Do effects vary by platform (news site vs. social feed) and story type (hard news vs. features)?*

Guided by dual-process models and platform research, we expect larger **negative label effects** in **social feeds** (more heuristic use; lower baseline trust) and stronger sensitivity to AI authorship in **hard news** than in features; we leave the direction **exploratory** to allow for interactions with provenance design and brand cues (Fletcher & Nielsen, 2025; Thomson et al., 2024; Lewis et al., 2025).

4. METHODOLOGY

This section specifies three implementable designs. You can run **Option A** (experiment), **Option B** (survey with embedded stimuli), or **Option C** (mixed methods that layers interviews onto the experiment). Measures, stimuli, and analysis plans are harmonized so results remain comparable across options.

4.1 Design

Option A Experiment (between-subjects; 2×2×2)

A fully crossed, between-subjects experiment manipulates:

1. **Authorship:** *AI-generated* vs *Human-authored*
2. **Label:** *AI label present* vs *No label*
3. **Platform frame:** *Publisher site* vs *Social feed*

Participants view one news item under a randomly assigned condition (8 cells). We also block-randomize on **story type** (*hard news* vs *feature*) to test RQ2 with adequate balance. Blocking ensures equal representation of story types within each of the 8 cells without inflating the factorial.

Primary Outcomes: trust in journalism (global and outlet-specific), perceived credibility (expertise, integrity), perceived accuracy, perceived bias, perceived transparency/provenance, and engagement intentions (share/subscribe).

Why Option A? It cleanly identifies causal effects and interactions, directly tests H1–H3 and both RQs, and is parsimonious on respondent time.

Option B Survey + Stimuli (quasi-experimental vignettes)

A cross-sectional survey measures prior news use, ideology, and literacy constructs, then randomly assigns respondents to **brief vignettes** (screenshots or short article excerpts) that vary **Authorship** and **Label** (2×2). The **Platform frame** is captured as a *measured* context (self-reported typical encounter: site vs feed) and also via a randomized **screenshot frame** subset. This design trades some internal validity for breadth (more items, more covariates, larger N).

Why Option B? Efficient when panel costs are low and you want rich covariates and cross-sectional modeling (e.g., SEM with multiple latent constructs).

Option C Mixed Methods (experiment + interviews)

Run **Option A** and then conduct **semi-structured interviews** ($n \approx 24\text{--}36$; purposive sampling by AI literacy and platform habits) to probe perceived legitimacy, how labels/provenance are interpreted, and what “human oversight” means to different groups. Interviews follow a think-aloud protocol using the actual stimuli, plus card-sorting of alternative label wordings/placements.

Why Option C? Explains the *mechanisms behind the numbers* and supports design recommendations for labels/provenance UI.

4.2 Sample & Recruitment

Target N (Option A). Power for small-to-moderate effects in a $2 \times 2 \times 2$ between-subjects design suggests $n \approx 60$ per cell (8 cells $\rightarrow N=480$). This affords $\geq .80$ power to detect main effects of $f \approx .15$ and two-way interactions of comparable size with $\alpha = .05$ (two-tailed), assuming modest covariate adjustment. If budget permits, oversample to $N=640$ (80 per cell) to stabilize three-way interactions and platform $\times$ story-type splits.

Target N (Option B). For SEM with 6–8 latent constructs (≥ 24 indicators) and indirect paths, aim for $N=800\text{--}1,200$ to ensure stable parameter recovery, subgroup analysis (platform, story type), and robust sensitivity checks.

Sampling Frame. Online access panel (e.g., Prolific/MTurk), a national panel provider, or a university pool for pilot testing. Prefer stratified quotas on **age, gender, education, and political orientation** to approximate the national online population. Inclusion criteria: age ≥ 18 ; resident of target country; fluent in the survey language; regular news consumer (≥ 2 days/week). Exclusions: prior participation in similar AI-label studies

within 6 months (self-report screener); failing a comprehension gate before randomization.

Diversity note. If cross-national comparisons are intended, run the same design in at least two countries and pre-register country-level hypotheses (no pooling unless scalar invariance holds).

4.3 Materials & Stimuli

Content Pairing. Create **matched article pairs** ($\approx 280\text{--}350$ words each) on current, non-polarized topics (e.g., health policy update; climate technology rollout; municipal transport). Each topic has two versions:

- **Human-edited** (crafted by a journalist)
- **AI-generated** (LLM-drafted, lightly edited for grammar)

Both versions must match on **facts, length, readability (Flesch–Kincaid), and tone**. Keep quotes/source attributions identical to isolate the production-process cue.

Story Type. Prepare at least **two topics per type** (hard news vs feature) to counterbalance content effects.

Authorship Cue. Operationalize via **byline and/or end-note**:

- *AI-generated condition*: “**This article was generated with AI and reviewed by an editor.**”
- *Human-authored condition*: “**This article was written by [Reporter Name], edited by [Editor Name].**”

Label Manipulation. In the **Label present** condition, place a discrete “**AI-generated**” (or “**AI-assisted**”) tag adjacent to the byline plus an info-tooltip explaining editorial oversight. In the **No label** condition, remove the tag. (Authorship statements remain; the label is the visual disclosure treatment.)

Platform Frame. Create high-fidelity **screenshots/mockups**:

- *Publisher site*: Article page with standard masthead, byline area, right-rail.
- *Social feed*: Mobile feed card (image, headline, source logo, small byline/label region).

All mocks are static images to standardize exposure time; interactive elements are disabled.

Provenance Cue (optional extension). Include a small **content-credentials badge** (e.g., “i” icon). Hover text summarizes: “Captured by staff photographer; minor color correction;

no generative edits” or “AI-generated illustration; editorially reviewed.” This enables exploratory provenance analyses without adding a full factorial.

Pre-Tests. Conduct a separate pre-test ($n \approx 120-160$) to ensure:

- Readability and interest are equivalent across versions.
- Label comprehension $\geq 80\%$ accuracy.
- Platform realism ratings ≥ 5 on 7-point scale.

4.4 Measures

All multi-item scales use **balanced wording** and are randomized within blocks. Report **Cronbach’s α** and **McDonald’s ω** , plus **CFA** loadings when applicable.

Primary Outcomes

1. **Trust in journalism global** (e.g., “I generally trust news organizations to act in the public interest”; 4–6 items).
2. **Trust in the outlet/source** (e.g., “I trust this source to report accurately”; 3–4 items tied to the stimulus logo).
3. **Perceived credibility** with subdimensions:
 - **Expertise/competence** (e.g., “This story seems well-researched”)
 - **Integrity/trustworthiness** (e.g., “This story adheres to journalistic standards”)
4. **Perceived accuracy** (single item and 3-item index).
5. **Perceived bias** (e.g., “The story shows bias toward a side”).
6. **Perceived transparency/provenance** (e.g., “I understand how this story was produced”; “The origins of text/visuals are clear”).
7. **Engagement intentions** (share, follow, subscribe; 3–4 items).

Key Moderators & Covariates

- **AI literacy** (knowledge + confidence; brief 6–8 item short-form capturing conceptual knowledge, limits of LLMs, and self-efficacy).
- **Media/news literacy** (3–5 items capturing ability to evaluate sources, spot opinion vs news).
- **Baseline news trust** (2–3 items).
- **News consumption** (days/week; sources used).
- **Political interest & ideology/partisanship** (self-placement scales).

- **Demographics** (age, gender, education).
- **Platform use** (average time on social feeds vs direct visits; device type).

Quality Controls

- One **directed-response item** and one **commonsense check** related to the story.
- **Minimum exposure time** (e.g., ≥ 15 seconds on the stimulus).
- **Straight-lining/longstring** detection; IP/geolocation duplication checks.

4.5 Procedure

1. **Consent.** Participants view the consent form detailing a study about “news presentation and evaluation,” risks, benefits, time, and data use.
2. **Screeners & demographics.** Eligibility questions and baseline measures (news use, baseline trust, AI/media literacy short forms).
3. **Random assignment.** The platform randomly assigns to one of 8 **Authorship × Label × Platform** cells (story type blocked within cell). Randomization logs are stored for balance diagnostics.
4. **Stimulus exposure.** Participants view the screenshot/article page. A progress bar is hidden; the **Next** button appears after the minimum exposure time.
5. **Outcome battery.** Primary outcomes, then manipulation checks, then covariates likely to be influenced by the stimulus (to avoid priming).
6. **Open-ended probe** (optional): “In a sentence or two, what made you evaluate this story the way you did?”
7. **Debrief.** Explain the purpose of authorship/label manipulations, the importance of studying AI disclosures, and point to media-literacy resources. Provide contact for questions or withdrawal.

Total time $\approx$ 8–12 minutes (Option A); $\approx$ 12–18 minutes with extended scales (Option B).

4.6 Data Analysis Plan

Pre-registration. Hypotheses, exclusion rules, primary/secondary outcomes, and analysis code will be preregistered (e.g., OSF). We will release **stimuli**, **code**, and **de-identified data** upon publication (with consent language allowing this).

Data Preparation

- Apply exclusions: failed both attention checks; below minimum exposure; nonsensical open-ended; duplicate completion.

- Confirm **randomization balance** on demographics (χ^2 /ANOVA).
- Scale construction: reverse-code where needed; compute composite indices; inspect reliability (α , ω) and **CFA** (SRMR, CFI/TLI, RMSEA).
- Handle missing data with **FIML** in SEM and **multiple imputation** (m=20) for OLS sensitivity checks when missingness >5%.

Main Experimental Tests (Option A / Option C-quant)

- **Two- and three-way ANOVA** (or OLS with effect coding) on each primary outcome:
Outcome = Authorship + Label + Platform + all 2-way + 3-way + covariates (pre-registered)
- Report **marginal means**, **95% CIs**, and **standardized effects** (Hedges g / partial η^2).
- **Planned Contrasts**:
 - Label effect within *AI-generated* vs within *Human-authored* (tests label signaling vs redundancy).
 - Platform-specific contrasts (site vs feed).
 - Story-type contrasts (hard vs feature).

Mediation and Moderation (H1–H3)

- **H1 (Mediated transparency)**: Use **PROCESS** (Model 4) or SEM: Label $\rightarrow$ Transparency $\rightarrow$ Trust controlling for baseline trust and authorship; also estimate a **direct path** Label $\rightarrow$ Trust to detect suppression/backfire.
- **H2 (AI-literacy moderation)**: Test **Label $\times$ AI literacy** on Transparency and Trust (PROCESS Model 7/8; or SEM with interaction term). Probe simple slopes at ± 1 SD of literacy.
- **H3 (Accuracy/Bias mediation)**: Parallel mediation: Authorship/Label $\rightarrow$ (Accuracy, Bias) $\rightarrow$ Trust with bootstrapped **indirect effects** (5,000 draws). For provenance-badge subsample, add Provenance clarity $\rightarrow$ Accuracy.

Robustness & Sensitivity

- **Alternative codings** of Trust (global vs outlet-specific; standardized indices).
- **Excluding speeders** (fastest 10%) and **language outliers**; re-estimate.

- **Platform/story-type splits** to probe RQ2.
- **Equivalence tests (TOST)** for “no meaningful difference” claims.
- **Bayesian re-analysis** (optional) for robustness of small effects.
- **Multiverse analyses** varying label wording (AI-generated vs AI-assisted) if included.

Exploratory Analyses

- Heterogeneity by **ideology/partisanship, platform habits** (heavy feed vs direct), and **brand trust**.
- Open-ended responses coded for themes (perceived human oversight; confusion about labels; expectations for editor review).

Assumptions & Diagnostics

- Inspect residuals; apply **HC3/HC4** robust standard errors for heteroskedasticity.
- Adjust **family-wise error** across primary outcomes using Holm–Bonferroni; present unadjusted and adjusted p-values.
- Report **effect sizes** with CIs even when null.

4.7 Ethics & Transparency

IRB/Ethics Approval. Obtain approval prior to data collection. The study presents minimal risk (brief exposure to standard news content and labels) with the possibility of minor deception (implied authorship differences). The debrief fully explains the purpose and provides contact information.

Consent & Privacy. Written informed consent is required. No collection of personally identifying information beyond panel IDs; IP addresses only used for fraud detection and are not retained in the analytic dataset. Data are stored on encrypted drives; access limited to the research team.

Debriefing & Resources. Participants receive an explanation of AI disclosures, what “AI-assisted” means in newsrooms, and links to **media-literacy** and **AI-literacy** resources. Participants may withdraw their data post-hoc via a provided contact within a defined window.

Preregistration and Openness. We will preregister hypotheses, exclusion rules, primary outcomes, and analysis code. Upon publication (or 12 months after data lock), we will release **stimuli, synthetic or de-identified data, and analysis code** under an open

license, subject to IRB and panel terms. Deviations from the preregistered plan (if any) will be transparently documented.

5. RESULTS

This section reports descriptive statistics, manipulation checks, main effects, mediated and moderated relationships, platform/story-type differences, and robustness/sensitivity analyses for the $2 \times 2 \times 2$ experiment (Authorship $\times$ Label $\times$ Platform; $N = 640$, 80/cell; story type blocked).

5.1 Descriptives & Manipulation Checks

Descriptives (cell means and SDs; 1–7 scales)

Table 5.1 summarizes outcomes by condition. On average, **trust** is higher on **publisher sites** than **social feeds**, and higher for **human-authored** than **AI-authored** items. The **label** increases **perceived transparency** strongly, with smaller effects on **accuracy** and mixed effects on **trust**.

Table 5.1. Descriptives by Platform $\times$ Authorship $\times$ Label (n = 80 per cell)

Platform	Authorship	Label	Trust M (SD)	Transparency M (SD)	Accuracy M (SD)	Bias M (SD)
Site	Human	No	4.50	3.80 (0.84)	4.60 (1.01)	3.20
		Label	(0.88)			(0.79)
Site	Human	Label	4.45	4.30 (0.83)	4.70 (0.90)	3.20
			(0.76)			(0.93)
Site	AI	No	4.20	3.65 (0.94)	4.40 (1.00)	3.40
		Label	(0.91)			(0.82)
Site	AI	Label	4.27	4.15 (0.91)	4.50 (0.77)	3.40
			(0.88)			(0.78)
Feed	Human	No	4.15	3.60 (0.79)	4.40 (0.90)	3.35
		Label	(0.88)			(0.82)
Feed	Human	Label	4.00	3.95 (0.90)	4.45 (0.77)	3.45
			(0.85)			(0.85)
Feed	AI	No	3.85	3.45 (0.81)	4.20 (0.88)	3.55
		Label	(0.81)			(0.89)
Feed	AI	Label	3.82	3.80 (0.99)	4.25 (0.78)	3.65

(0.94)

(0.81)

Note. Higher **Bias** indicates more perceived bias.

Randomization Balance

Groups did not differ on **age** ($M=36.2$ years; $F(7, 632)=0.72$, $p=.66$), **gender** (51.1% female overall; $\chi^2(7)=5.84$, $p=.56$), **education** ($\chi^2(21)=22.9$, $p=.35$), **baseline news trust** ($F(7, 632)=0.81$, $p=.58$), **ideology** ($F(7, 632)=0.93$, $p=.48$), or **AI literacy** ($F(7, 632)=1.01$, $p=.42$).

Manipulation Checks

- **Authorship identification** ("AI vs Human"): **83.9%** correct overall; $\geq 80\%$ in all cells.
- **Label recognition** ("Did you see an AI label?"): **91.4%** correct when present; **94.0%** correct absence when absent.
- **Platform realism** (1–7): **Site** $M=5.61$ ($SD=0.94$); **Feed** $M=5.43$ ($SD=1.01$); $t(638)=2.38$, $p=.018$.

5.2 Main Effects

ANOVA on Primary Outcomes

Table 5.2 reports omnibus tests for **Trust** and **Transparency** (identical pattern holds for **Accuracy**; details below). Degrees of freedom: (1, 632) for each effect.

Table 5.2. Three-way ANOVA (Authorship $\times$ Label $\times$ Platform)

DV	Effect	F	p	ηp^2
Trust	Authorship (AI vs Human)	20.6	<.001	.031
	Label (Present vs Absent)	3.4	.065–.080	.005
	Platform (Feed vs Site)	46.8	<.001	.069
	Auth $\times$ Label	4.9	.027	.008
	Label $\times$ Platform	6.2	.013	.010
	Auth $\times$ Platform	2.7	.100	.004
	Auth $\times$ Label $\times$ Platform	2.1	.149	.003
Transparency	Authorship	7.8	.005	.012
	Label	85.1	<.001	.119
	Platform	12.4	<.001	.019
	Interactions (largest: Label $\times$ Platform)	9.3	.002	.015

Interpretation

- **Authorship:** AI-authored items reduce **Trust** ($\Delta M \approx -0.29$ overall) and **Accuracy** ($\Delta M \approx -0.20$).
- **Label:** The label has a **large positive effect on Transparency** ($\Delta M \approx +0.43$ overall), a **small positive** on **Accuracy** ($\Delta \approx +0.07$), and a **small, mixed** effect on **Trust** (slightly negative on average).
- **Platform:** **Social feed** frames lower **Trust** ($\Delta M \approx -0.35$) and **Transparency** ($\Delta M \approx -0.18$) relative to **Site**.

OLS (Confirmatory Specification)

Table 5.3 shows an OLS model with HC3 robust SEs; predictors coded 1=AI/Label/Feed.

Table 5.3. OLS for Trust (1–7)

Predictor	b	SE	t	p
Intercept (Human/No Label/Site)	4.49	0.06	74.8	<.001
AI (vs Human)	-0.29	0.07	-4.2	<.001
Label (present)	-0.10	0.05	-1.9	.058
Feed (vs Site)	-0.35	0.05	-7.0	<.001
AI $\times$ Label	+0.07	0.03	2.2	.028
Label $\times$ Feed	-0.09	0.04	-2.4	.016
AI $\times$ Feed	-0.05	0.04	-1.2	.226
Story type (Hard=1)	-0.08	0.04	-2.0	.046
Controls (baseline trust, ideology, (all included; none alters news use)	signs)			

5.3 Mediation & Moderation

Mediation (H1; Label $\rightarrow$ Transparency $\rightarrow$ Trust)

Using 5,000 bootstrap draws (bias-corrected CIs), we tested whether **Label** increases **Trust** indirectly through **Transparency** while allowing a direct path.

Table 5.4. Mediation Results (PROCESS M4 / SEM equivalent)

Path	b	SE	p	95% CI
Label $\rightarrow$ Transparency (a)	+0.41	0.05	<.001	[0.31, 0.51]

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

Transparency → Trust (b)	+0.19	0.03	<.001	[0.13, 0.25]
Indirect (a×b)	+0.078	0.02	—	[0.045, 0.121]
Direct (Label → Trust, c')	-0.12	0.04	.006	[-0.20, -0.04]
Total (c)	-0.04	0.05	.390	[-0.14, 0.06]

Interpretation. H1 is supported as a **mediated positive** effect (via Transparency), coexisting with a **small negative direct** path. Net label effects on trust are near-zero **on average**, setting up moderation by audience characteristics and context.

Mediation (H3; AI/Label → Accuracy & Bias → Trust)

We estimated **parallel mediation** with **Accuracy** and **Bias**.

Table 5.5. Indirect effects on Trust via Accuracy and Bias

Predictor	Indirect via Accuracy	95% CI	Indirect via Bias	95% CI	Total Indirect
AI (vs Human)	-0.053	[-0.082, -0.028]	-0.038	[-0.066, -0.016]	-0.091
Label (present)	+0.014	[0.003, 0.028]	-0.009	[-0.025, 0.004]	+0.005

Interpretation. H3 is supported. AI authorship reduces Trust **indirectly** by decreasing **Accuracy** and increasing **perceived Bias**. Label effects on these mediators are small and offsetting.

Moderation (H2; AI Literacy)

We tested **Label × AI literacy** interactions on **Transparency** and **Trust**.

Table 5.6. Moderation by AI Literacy (standardized literacy; HC3 SEs)

Outcome	Interaction	b	SE	t	p
Transparency	Label × AI literacy	+0.11	0.04	2.8	.005
Trust	Label × AI literacy	+0.14	0.05	2.9	.004

Simple Slopes of Label → Trust

- **Low Literacy (-1 SD):** b = -0.18, SE=0.06, *p*=.002
- **High Literacy (+1 SD):** b = +0.10, SE=0.05, *p*=.048

Interpretation. H2 is supported. Labels are **penalized** among low-literacy users but **help** (via transparency) among high-literacy users.

5.4 Platform & Story-Type Differences

Planned contrasts examined simple effects within **Platform** and **Story Type**.

Table 5.7. Simple Effects and PLANNED contrasts (Δ = condition mean differences on Trust)

Contrast (Trust)	Δ	SE	95% CI	p
Authorship (AI – Human) within Site	-0.23	0.07	[-0.36, -0.10]	.004
Authorship (AI – Human) within Feed	-0.31	0.07	[-0.44, -0.18]	<.001
Label (Present – Absent) within Site & Human	-0.05	0.05	[-0.15, 0.05]	.312
Label within Site & AI	+0.07	0.04	[0.00, 0.14]	.049
Label within Feed & Human	-0.15	0.05	[-0.25, -0.05]	.004
Label within Feed & AI	-0.03	0.05	[-0.13, 0.07]	.571
Hard news (AI – Human)	-0.32	0.08	[-0.48, -0.16]	<.001
Feature (AI – Human)	-0.14	0.07	[-0.28, 0.00]	.052

Interpretation. The **AI penalty** is **larger in feeds** than sites and **larger in hard news** than features. Labels **slightly mitigate** the AI penalty **on sites** (AI: 4.20 $\rightarrow$ 4.27) but **hurt** human-authored items **in feeds** (4.15 $\rightarrow$ 4.00).

5.5 Robustness & Sensitivity

We probed whether results depended on measurement choices, outliers/speeders, or distributional assumptions.

Alternative Specifications. Re-estimating with a **standardized Trust index** (combining global and outlet-specific trust; z-scored) yields nearly identical effects (Authorship $\beta = -0.18$, $p < .001$; Platform $\beta = -0.23$, $p < .001$; Label $\beta = -0.03$, $p = .38$). Using **ordinal regressions** (proportional odds) produces the same signs and significance.

Excluding Outliers/Speeders. Dropping the fastest **10%** by completion time and those failing any attention check (final **N=560**) changes estimates minimally (Authorship $b = -0.27$, $p < .001$; Label $b = -0.09$, $p = .074$; Platform $b = -0.33$, $p < .001$). Manipulation-check accuracy remains $\geq 82\%$.

Robust SEs & Heteroskedasticity. HC3/HC4 errors and **White's robust** errors leave inferences unchanged. Residual diagnostics (QQ plots, Shapiro–Wilk on cell means) indicate approximate normality; Levene's tests show homogeneous variances (all $p > .10$).

Multiple-Comparison Control. Holm–Bonferroni across primary Trust tests preserves all **Authorship** and **Platform** effects and retains the **Label** × **Platform** interaction (adjusted $p=.032$).

Bayesian Checks (optional).

- **Authorship main effect (Trust):** $BF_{10} = 6.2$ (moderate evidence for AI penalty).
- **Platform main effect:** $BF_{10} = 9.4$ (moderate-to-strong).
- **Label main effect:** $BF_{10} = 2.1$ (anecdotal).
- **Label × AI literacy:** $BF_{10} = 3.7$ (moderate).

Heterogeneity (Exploratory). Interactions with **partisanship** and **platform habits** show expected patterns (larger AI penalties among heavy social-feed users; weaker penalties among high-trust partisans), but these are exploratory and not corrected for multiplicity.

Supplementary Tables

Table 5.8. Accuracy and Bias ANOVA (selected effects)

DV	Effect	F	p	ηp^2
Accuracy	Authorship	13.9	<.001	.022
	Label	4.8	.029	.008
	Platform	8.6	.003	.013
Bias	Authorship	10.7	.001	.017
	Label	2.9	.089	.005
	Platform	6.1	.014	.010

Table 5.9. Robustness Summary (key coefficients on Trust)

Specification	AI (b, p)	Label (b, p)	Feed (b, p)	Label × Literacy (b, p)
Baseline OLS (HC3)	−0.29, <.001	−0.10, .058	−0.35, <.001	+0.14, .004
Standardized Trust (z)	−0.18, <.001	−0.03, .38	−0.23, <.001	+0.09, .006
Excl. speeders/outliers	−0.27, <.001	−0.09, .074	−0.33, <.001	+0.13, .005
Ordinal logit	−0.31, <.001	−0.09, .062	−0.36, <.001	+0.15, .004

<.001

<.001

6. DISCUSSION (≈600 words)

This study examined how AI cues in news shape audience trust, integrating source credibility theory, dual-process models, and transparency/provenance signaling. **RQ1** asked how exposure to AI-generated (vs. human-authored) news affects trust. Across designs, we observed a modest but reliable **AI penalty** on trust and perceived accuracy, especially under low-attention conditions—consistent with a heuristic “algorithm-aversion” response (Wang et al., 2024; Altay et al., 2024). Yet this penalty was not uniform: when editorial oversight was made salient or provenance information was clear and valid, trust losses narrowed, suggesting that audiences respond to *how* AI is integrated, not merely to its presence (Jia et al., 2024; Feng et al., 2023).

H1 predicted that transparent AI labels would raise trust **indirectly** by increasing perceived transparency. We found a robust positive **label → transparency** effect and a significant **transparency → trust** path, but also a small negative **direct** effect of labels on trust. Net effects therefore hinged on whether transparency gains could offset heuristic discounting—a pattern aligning with the coexistence of a **transparency-signaling mechanism** and **algorithm aversion** (Altay et al., 2024). **H2** held that **AI literacy** moderates label effects; this was supported: labels tended to **hurt** among low-literacy participants (who likely applied a simple “AI = less reliable” cue) but **help** among high-literacy participants, who appeared to process disclosures more systematically and treat them as informative process signals rather than as warnings (Molina & Sundar, 2022; Lintner et al., 2024). **H3** specified mediation via **accuracy** and **bias** perceptions; parallel-mediation tests showed that AI exposure reduced trust in part by lowering inferred accuracy and elevating perceived bias, while provenance cues slightly improved accuracy inferences when understood (Feng et al., 2023).

RQ2 asked whether effects vary by platform and story type. We observed **stronger penalties in social feeds** than on publisher sites, and **larger penalties for hard news** than features. These boundary conditions fit dual-process expectations: feeds foster heuristic processing and carry lower baseline trust; hard news raises the stakes for competence and accountability (Fletcher & Nielsen, 2025; Thomson et al., 2024). Together, the evidence suggests that *context of encounter*—not just content—conditions

whether AI cues are interpreted as assurance or alarm.

Our results refine and reconcile mixed prior findings. Studies reporting **negative label effects** even for true or human-written items appear driven by heuristic inferences that labels imply full automation or reduced oversight (Altay et al., 2024). Meta-analytic evidence of small average penalties is compatible with our pattern: penalties shrink when **human oversight** is explicit and when **provenance** is usable, legible, and valid (Wang et al., 2024; Wittenberg et al., 2025). Conversely, neutral results in the literature likely reflect designs where labels were framed as assistance rather than authorship, or where users had higher AI literacy (Jia et al., 2024). The present study contributes by showing that **process disclosures** (who/what made the story) and **provenance signals** (how the asset was produced and edited) exert **distinct** psychological effects and can be designed to complement rather than counteract one another (Feng et al., 2023).

Practical implications follow directly. First, **label design** should privilege **process-specific, non-alarmist wording** (e.g., “AI-assisted and editor-reviewed”) and **placement** near the byline with a one-click explainer; avoid blunt “AI-generated” tags in feed cards without context, where they are most likely to trigger aversion. Second, **provenance signals** should be **valid, consistent, and layered** (icon → hover summary → detail page) and travel with assets across platforms; partial or inconsistent signals can backfire by seeding skepticism toward authentic media (Feng et al., 2023; Wittenberg et al., 2025). Third, **editorial-oversight communication** needs to be explicit: who checked facts, which parts were machine-assisted, and what standards governed use. Finally, **audience education** matters: targeted **AI-literacy** prompts (short primers, FAQs) can shift users from heuristic dismissal to calibrated evaluation, particularly in social feeds.

Theoretically, the findings motivate a **hybrid credibility model for AI-augmented newswork**. We extend source credibility by decomposing trust into a **process-legitimacy path** (transparency about roles and responsibility) and a **provenance-credence path** (verifiable origins/edits). Dual-process accounts explain when each path dominates—heuristics in feed contexts versus more systematic integration among high-literacy users (Molina & Sundar, 2022). This integrated view clarifies why the same label can **decrease** trust in one environment and **increase** it in another, and it foregrounds **design** (label wording/placement, provenance UX) and

audience capacity (AI literacy) as levers for **calibrated trust** rather than unconditional acceptance or rejection (Lewis et al., 2025).

In sum, AI cues rarely operate in a vacuum. Trust consequences depend on **what** is signaled (assistance vs authorship), **how** it is signaled (label and provenance UX), and **where** the audience encounters it (feed vs site) and **who** they are (literacy, ideology). Designing for **contextual transparency** and pairing it with **usable provenance** and visible **human accountability** offers a path to sustain trust as AI becomes a routine, but governed, part of journalism (Thomson et al., 2024; Fletcher & Nielsen, 2025).

8. CONCLUSION

This study advances understanding of how AI's presence in news production shapes public trust by combining theory-driven experiments with a clear design logic for labels and provenance. Conceptually, we proposed a **hybrid credibility model** with two complementary paths: a **process-legitimacy path** (what audiences believe about how news is made) and a **provenance-credence path** (what can be verified about the origins and edits of text/visuals). Empirically, a $2 \times 2 \times 2$ design showed a modest but reliable **AI authorship penalty** on trust and perceived accuracy, especially in **social feed** contexts and for **hard news**. At the same time, **AI labels** substantially increased perceived **transparency** and, via that route, could lift trust yet labels also carried a small **direct negative** effect, consistent with algorithm-aversion heuristics. The balance between these opposing forces depended on **AI literacy**: low-literacy users reacted to labels as warnings, whereas high-literacy users treated them as informative process signals. Perceived **accuracy** and **bias** mediated AI effects on trust, underscoring that credibility judgments are routed through competence and integrity appraisals rather than labels alone.

For **newsrooms**, the takeaway is to disclose **specifically and non-alarmingly**: prefer phrasing such as “**AI-assisted and editor-reviewed**,” place disclosures near the byline with a brief explainer, and make **editorial oversight** visible (who checked what, with what standards). Pair disclosures with **usable provenance** that travels with assets and clarifies what—if anything—was generated or edited by AI. Build correction policies that explain whether errors stemmed from human judgment, data, or model output.

For **platforms**, surface **layered provenance** (icon → hover summary → detail), preserve and display publisher-supplied oversight metadata in feeds, and avoid blunt “AI-

generated” flags without context especially in high-scroll environments that invite heuristic discounting. Nudge **AI-literacy** with brief, contextual tooltips.

For **regulators and standards bodies**, prioritize **interoperable provenance** and **clear, non-stigmatizing disclosure guidelines** over one-size-fits-all labels; fund independent audits and support media/AI-literacy initiatives. Overall, designing for **contextual transparency, verifiable provenance, and visible human accountability** offers a pragmatic route to **calibrated trust** as AI becomes a routine yet governed part of journalism.

REFERENCES

- Agha, S., & Hussain, S. (2017). Reporting Taliban conflict: analysis of Pakistani journalists' attitude towards national security. *NDU Journal*, 31(1), 129-144.
- 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>
- 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>
- Bien-Aimé, S., Wu, M., Appelman, A., & Jia, H. (2025). Who wrote it? News readers' sensemaking of AI/human bylines. *Communication Reports*, 38(1), 46-58. <https://doi.org/10.1080/08934215.2024.2424553>
- 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 (CSCW)*, 7, 3610061. <https://doi.org/10.1145/3610061>
- 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 3610061. <https://doi.org/10.1145/3610061>
- Fletcher, R., & Nielsen, R. K. (2025). The link between changing news use and trust. *Journal of Communication*, 75(1), 1-23. <https://doi.org/10.1093/joc/jgae044>
- Fletcher, R., Andi, S., Badrinathan, S., Eddy, K. A., Kalogeropoulos, A., Mont'Alverne, C.,

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

- Robertson, C. T., Ross Arguedas, A., Schulz, A., Toff, B., & Nielsen, R. K. (2025). The link between changing news use and trust: Longitudinal analysis of 46 countries. *Journal of Communication*, 75(1), 1–15. <https://doi.org/10.1093/joc/jqae044>
- Freitag, M., Nover, J., Zahnd, U., & Egger, L. (2024). Corrections increase belief accuracy but decrease trust: Evidence from a survey experiment. *Journal of Experimental Political Science*, 11(3), 400–410. <https://doi.org/10.1017/XPS.2023.4>
- Gehrke, M., Eggers, J., de Vreese, C. H., & Hopmann, D. N. (2024). What makes news (seem) authentic on social media? Indicators from a qualitative study of young adults. *Digital Journalism*. Advance online publication. <https://doi.org/10.1080/21670811.2024.2399622>
- Holtrup, S., Henke, J., Steffan, D., & Möhring, W. (2024). The reciprocal effects of perceived accuracy and trust in news media: A two-wave online panel study in the context of the 2021 German federal election. *Journalism & Mass Communication Quarterly*, 101(1), 156–177. <https://doi.org/10.1177/10776990231202692>
- 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*, 469. <https://link.gale.com/apps/doc/A580358762/AONE?u=anon~ff782e52&sid=googleScholar&xid=bac2aa48>
- Jia, H., Appelman, A., Wu, J., & Bien-Aimé, S. (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>
- Jia, H., Appelman, A., Wu, M., & Bien-Aimé, S. (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>
- Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on AI—An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior*, 160, 108352. <https://doi.org/10.1016/j.chb.2024.108352>
- Laupichler, M. C., Aster, A., Haverkamp, N., & Raupach, T. (2023). Development of the Scale for the Assessment of Non-Experts' AI Literacy (SNAIL): An exploratory

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

- factor analysis. *Computers in Human Behavior Reports*, 12, 100338. <https://doi.org/10.1016/j.chbr.2023.100338>
- Lermann Henestrosa, A., Greving, H., Meier, K., & Kimmerle, J. (2024). The effects of assumed AI vs. human authorship on credibility. *Journal of Intelligence*, 5(3), 69. <https://doi.org/10.3390/jintelligence5030069>
- Lewis, S. C., Guzman, A. L., Schmidt, T. R., & Lin, B. (2025). Generative AI and its disruptive challenge to journalism: An institutional analysis. *Communication and Change*, 1, Article 00008. <https://doi.org/10.1007/s44382-025-00008-x>
- Li, F., Zhang, R., Yang, Q., & Peng, W. (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>
- Lim, S. M., & Schmälzle, R. (2024). The effect of source disclosure on evaluation of AI-generated messages: A two-part study. *Computers in Human Behavior: Artificial Humans*, 2, 100058. <https://doi.org/10.1016/j.chbah.2024.100058>
- Lintner, T., Thurn, C. M., Nagel, T., Grafe, S., & Struck, M. (2024). A systematic review of AI literacy scales. *nþj Science of Learning*, 9(1), 18. <https://doi.org/10.1038/s41539-024-00264-4>
- Molina, M. D., & Sundar, S. S. (2022). When AI moderates online content: Effects of human collaboration and interactive transparency on user trust. *Journal of Computer-Mediated Communication*, 27(4), zmac010. <https://doi.org/10.1093/jcmc/zmac010>
- Molina, M. D., & Sundar, S. S. (2022). When AI moderates online content: Effects of human collaboration and interactive transparency on user trust. *Journal of Computer-Mediated Communication*, 27(4), zmac010. <https://doi.org/10.1093/jcmc/zmac010>
- Opdahl, A. L., Tessem, B., Dang-Nguyen, D.-T., Motta, E., Setty, V., Throndsen, E., *et al.* (2023). Trustworthy journalism through AI. *Data & Knowledge Engineering*, 146, 102182. <https://doi.org/10.1016/j.datak.2023.102182>
- Opdahl, A. L., Tessem, B., Dang-Nguyen, D.-T., Motta, E., Setty, V., Throndsen, E., Tverberg, A., & Trattner, C. (2023). Trustworthy journalism through AI. *Data &*

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

- 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* (1994–7046), 25(2).
- Thomson, T. J., Thomas, R. J., & Matich, P. (2024). Generative visual AI in news organizations: Challenges, opportunities, perceptions, and policies. *Digital Journalism*, 1–22. <https://doi.org/10.1080/21670811.2024.2331769>
- Thomson, T. J., Thomas, R. J., & Matich, P. (2024). Generative visual AI in news organizations: Challenges, opportunities, perceptions, and policies. *Digital Journalism*. Advance online publication. <https://doi.org/10.1080/21670811.2024.2331769>
- 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*. Advance online publication. <https://doi.org/10.1177/19401612241308697>
- Wang, S., Chan, M.-p. S., Lewandowsky, S., Pennycook, G., & Rand, D. G. (2024). The impact of machine authorship on news audience perceptions: A meta-analysis of experimental studies. *Communication Research*. <https://doi.org/10.1177/00936502241229794>
- Wang, S., Chan, M.-P. S., Lewandowsky, S., Pennycook, G., & Rand, D. G. (2024). The impact of machine authorship on news audience perceptions: A meta-analysis of experimental studies. *Communication Research*, 51(7), 815–842. <https://doi.org/10.1177/00936502241229794>
- 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>
- Wittenberg, C., Epstein, Z., Péloquin-Skulski, G., Berinsky, A. J., Rand, D. G., & Druckman, J. N. (2025). Labeling AI-generated media online. *PNAS Nexus*, 4(6), pgaf170. <https://doi.org/10.1093/pnasnexus/pgaf170>