Simen Owen Academic Proceedings Series

4th International Conference on Media, Economy, Communication and Intelligence Management (MECI 2026)

Article

The Moderating Role of Anthropomorphic Presentation in Audience Recognition of Generative AI News: A MAIN-Based Model and Experimental Protocol

Yunxing Lu[1,*], Shuqing Yang[1] and Mandi Xu[2]

[1] International College, Hebei University, Baoding, China
[2] College of Traditional Chinese Medicine, Hebei University, Baoding, China
* Correspondence: Yunxing Lu, International College, Hebei University, Baoding, China

Abstract: Humanlike news interfaces can make generative artificial intelligence appear socially familiar while leaving the origin of a news report uncertain. This article develops a conditional process model of audience recognition of AI authorship, drawing on the Modality, Agency, Interactivity, and Navigability model. It distinguishes recognition of production origin from perceived credibility, social presence, and identification with a presenter. Truthful authorship disclosure is specified as the predictor, anthropomorphic presentation as a randomized moderator, and source scrutiny and social presence as parallel mediators. A prospective factorial experiment crosses disclosure with presentation style and includes both AI-drafted and human-written news. The proposed analysis separates AI-attribution judgments from classification accuracy and examines false-positive judgments for human-written reports. It uses crossed participant–stimulus models for randomized treatment effects and participant means for explanatory conditional indirect associations. Competing predictions are retained where existing evidence does not justify a directional hypothesis: a disclosure may encourage source scrutiny or make further scrutiny unnecessary, while humanlike presentation may either compete with or heighten attention to provenance cues. The article provides operational definitions, a stimulus protocol, draft measures, and an analysis plan; it reports no new participant data or completed mediation test. Its contribution is a testable account of how interface design may condition the relationship between disclosure and recognition without treating social engagement as evidence of mistaken authorship attribution.

Keywords: generative AI news; anthropomorphic presentation; MAIN model; source recognition; moderated parallel mediation

Received: 07 September 2026
Revised: 23 October 2026
Accepted: 02 November 2026
Published: 05 November 2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Knowing who or what produced a news report is increasingly separate from judging whether its content is credible. An audience member can recognize AI involvement and still consider a report accurate, or accept a report as human-written while distrusting its claims. Early research on automated journalism already demonstrated that perceived credibility, readability, and production origin are distinct dimensions of audience evaluation [1,2]. Generative systems make this distinction especially consequential because their outputs can contain the conversational and narrative features that readers associate with human authorship.

Interface design adds a further source of uncertainty. A report may be introduced by a humanlike avatar, carry an explicit AI label, and appear within a professionally edited news service. These cues concern different actors and production stages. The avatar represents a presenter; the label can identify a drafting system; the news organization

assumes editorial responsibility. Treating all three as one source obscures the question audiences are being asked to answer. Experiments on machine attribution show that authorship labels and their placement can change news evaluations, but these effects do not establish whether readers correctly identify the underlying production process [3,4].

The MAIN model explains how technological affordances associated with modality, agency, interactivity, and navigability can activate heuristics used in evaluating mediated information [5]. Its agency dimension is particularly useful for examining competing cues about human and machine contribution. Nevertheless, the model was not formulated as a validated scale of AI detection ability. Extending it to authorship recognition requires an explicit account of the judgment target, independent evidence of actual authorship, and measures that distinguish recognition from trust. The broader human–AI interaction framework likewise places perceived machine agency within a relationship among technological features, user perceptions, and responses [6].

Anthropomorphic presentation is a plausible boundary condition rather than an inherently beneficial or harmful design choice. Humanlike cues can encourage social responses without creating a literal belief that an interface is human. Conversely, readers can use seemingly human linguistic features as unreliable evidence of human authorship. Research on AI-generated self-presentations demonstrates this problem, although its findings cannot be transferred directly to news without a domain-specific test [7]. The central issue is whether anthropomorphic presentation changes how audiences use reliable provenance information when making authorship judgments.

This article develops a model and prospective experiment addressing that issue. The predictor is the presence of a truthful authorship disclosure; the moderator is a separately manipulated anthropomorphic presentation; and the proposed mediators are source scrutiny and social presence. The primary judgment concerns whether generative AI drafted a report. Human-written comparison reports permit assessment of mistaken AI attribution as well as successful recognition. This design goes beyond a general acceptance model by specifying an observable production process and an outcome that can be checked against it.

Three research questions organize the contribution. First, does anthropomorphic presentation change the effect of disclosure on authorship judgments and classification performance? Second, are those conditional relationships consistent with parallel indirect associations through source scrutiny and social presence? Third, can changes in correct AI recognition be distinguished from a general tendency to classify all news as AI-generated? The article answers these questions at the level of theory and study design. Empirical support would require implementing the protocol and evaluating the resulting observations.

2. Literature Review and Hypothesis Development

2.1. Authorship Recognition as a Distinct Audience Outcome

Audience recognition is defined here as identifying the production origin of the report's written text. It does not mean identification with a media character, attachment to a news organization, or acceptance of AI journalism. Identification with media characters concerns adopting a represented character's perspective rather than classifying authorship [8]. Keeping the terms separate prevents a favorable attitude or a sense of connection from being interpreted as correct detection.

The source-monitoring framework provides a complementary explanation of origin judgments. People infer origins from features of an experience and from reflective evaluation; those inferences can be mistaken [9]. Applied to news interfaces, a face, personal name, or conversational style can become evidence about a presumed communicator even when it says little about who drafted the text. This application is a theoretical extension, not a claim that a specific source-memory experiment has already established an effect of AI news avatars.

The distinction between a source cue and the judgment it elicits is empirically important. Research comparing automated and human-written news has found context-dependent differences across credibility and readability assessments [10]. More recent work on news bylines distinguishes the displayed attribution from the AI contribution readers actually perceive; humanness perceptions help explain credibility judgments in that setting [11]. Neither result makes credibility a substitute for authorship accuracy. A report can receive a high AI-attribution score for the right or wrong reason, depending on its documented production origin.

Accordingly, the proposed study will collect a probability judgment and a categorical judgment. The first asks how likely it is that generative AI drafted the text. The second asks participants to classify the report as AI-drafted or human-written. Ground truth is defined by controlled production records. For AI-drafted reports, a higher AI probability indicates stronger correct attribution. For human-written reports, the same increase indicates stronger mistaken attribution. Accuracy and calibration must therefore be interpreted jointly across the two source categories.

2.2. MAIN and Competing Agency Cues

Within MAIN, agency cues concern the perceived source of communication, while modality concerns forms of presentation. Visual presentation can carry agency cues without changing the communication modality. An explicit AI-authorship label supplies information about the production agent; an anthropomorphic presenter supplies visual cues that may invite humanlike social interpretation. The proposed experiment varies those agency cues independently within a fixed visual modality, while holding navigation and interaction constant. It therefore examines a bounded application of MAIN rather than claiming to test all four affordances.

Anthropomorphism refers to attributing human characteristics to nonhuman entities. Its theoretical explanation includes the availability of human knowledge, motivations to understand an entity, and motivations for social connection [12]. It should not be equated with deception or with a single visual style. Research on computer anthropomorphism distinguishes explicit beliefs from social responses that can occur without a conscious belief that a system is human [13]. This distinction is central to the present model: social presence may persist alongside accurate awareness of AI authorship.

Direct news research also cautions against assuming that every favorable response is a credibility effect. Experiments on anthropomorphized algorithms have found changes in emotional involvement and liking, with effects depending on the relationship communicated by the interface [14]. These findings motivate attention to a social route but do not establish that humanlike presentation necessarily reduces recognition. Evidence from anthropomorphized chatbots further connects design cues with social presence in an adjacent communication context [15]. Transfer to news reading remains an empirical question.

The model treats disclosure and anthropomorphic presentation as potentially competing or complementary cues. A clear label might dominate a humanlike appearance. Alternatively, a socially engaging presenter might draw attention away from production information. Cue incongruity could also attract scrutiny: an apparently human presenter accompanied by an AI label may prompt closer inspection. These alternatives make the interaction theoretically informative even if its direction differs across stimuli or audiences.

2.3. Source Scrutiny and Social Presence as Parallel Mediators

Source scrutiny is defined as attention and reflective effort directed toward evidence about how a report was produced. It concerns examining provenance, not deciding that the text is AI-generated. Social presence is the felt sense of a socially present communicator during exposure. It concerns the experience of mediation, not belief in the presenter's biological existence. These definitions keep both mediators conceptually separate from the authorship outcome and from the experimental assignment.

The scrutiny route allows disclosure to change how production information is processed. Its direction is not predetermined. A conspicuous label could prompt further checking, but an unambiguous disclosure could also answer the authorship question immediately and reduce the need for additional scrutiny. Measuring only time spent inspecting the source would therefore miss cases in which recognition is accurate and effortless. Self-reported scrutiny and behavioral inspection will be analyzed separately, with their convergence treated as evidence to be assessed rather than assumed.

The social route allows humanlike presentation to affect the way disclosure is interpreted. A report may feel socially engaging even when its AI origin is understood. Disclosure could reduce that sense of presence by making the technological agent salient, or leave it intact because social responses do not require literal human attribution. The model consequently does not assume that every increase in social presence is accompanied by reduced recognition. A conditional association between presence and the authorship judgment is one empirical possibility among several.

The two mediators are specified in parallel to compare their distinct conditional associations within one model. Their residuals will be allowed to covary. The specification does not claim that scrutiny and social presence are independent psychological processes or that neither can influence the other. A serial model would require additional temporal evidence and a different identification strategy. The parallel formulation provides a transparent starting point while preserving the possibility that one pathway is absent or that their directions offset each other.

2.4. Hypotheses and Conditional Process Questions

H1: Truthful authorship disclosure will increase correct classification of production origin relative to a neutral publication tag, averaged across presentation conditions. On AI-drafted reports, it will increase AI-attribution probability; on human-written reports, it will decrease that probability.

H2: The anthropomorphic presentation condition will increase perceived social presence relative to the nonanthropomorphic condition, averaged across disclosure conditions. This is a hypothesis about the proposed interface manipulation, not an established effect for every humanlike design.

RQ1: Does anthropomorphic presentation moderate the effect of disclosure on AI-attribution judgments and classification performance?

RQ2: Does the indirect association between disclosure and AI attribution through source scrutiny differ between presentation conditions?

RQ3: Does the indirect association through social presence differ between presentation conditions, and do the two conditional indirect associations reinforce or offset each other?

The nondirectional questions reflect unresolved theoretical alternatives. Assigning a positive sign to every pathway would conceal the possibility that a clear disclosure improves recognition while reducing effort, or that social presence coexists with accurate AI awareness. The statistical plan will estimate the relevant interaction and indirect-association contrasts directly, following the distinction between moderation of a component path and moderation of its complete indirect association [16,17].

Figure 1. Conceptual path diagram linking authorship disclosure, anthropomorphic presentation, source scrutiny, social presence, and AI attribution judgment.

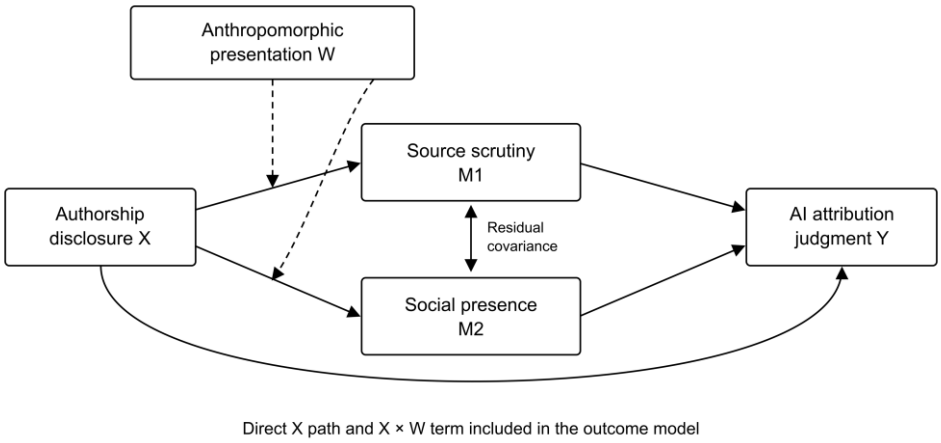


Figure 1. Proposed Model for AI-drafted Reports. Disclosure Is X, Anthropomorphic Presentation Is W, Source Scrutiny and Social Presence Are Parallel Mediators, and AI-attribution Probability Is Y. Solid Arrows Specify Regression Paths; Dashed Arrows Identify Moderation of the Two First-Stage Paths. The Model Also Includes X × W in the Outcome Equation and Permits Mediator Residual Covariance. Main Effects of W Are Included in All Three Regression Equations. No Path Coefficient Is an Empirical Estimate.

3. Methodology

3.1. Design and Research Setting

A prospective online experiment will recruit Mandarin-reading adults aged 18 years or older in mainland China. The design crosses two between-participant factors: authorship disclosure, present versus absent, and presentation, anthropomorphic versus nonanthropomorphic. Participants will be allocated with equal probability to the four substantive treatment cells. The measurement-order factor described below will be independently balanced within each cell, creating eight allocation combinations for recruitment and power planning. Both AI-drafted and human-written reports will appear in randomized order within each participant's session. Table 1 specifies the manipulations and the intended estimands.

Table 1. Experimental Factors and Source Benchmarks.

Component	Operationalization	Interpretation
Disclosure X	X = 1: truthful drafting-source label. X = 0: neutral publication tag in the same location.	Effect of providing provenance information, with layout held constant.
Presentation W	W = 1: humanlike presenter icon. W = 0: abstract presenter icon matched for size and visual prominence.	Effect of the specified visual presentation contrast.
Actual source S	AI-drafted and human-written reports, documented during stimulus production.	Separates correct AI recognition from false AI attribution.
Primary judgment Y	Probability from 0 to 100 that generative AI drafted the report.	Higher is more accurate only for AI-drafted reports.
Secondary performance	Classification accuracy, false-positive rate, Brier score, and calibration plots.	Evaluates performance across both actual sources.

The target construct is recognition of the text's drafting origin. All conditions will use the same display format and an interface presenter that is explicitly described in the general instructions as separate from the author. The humanlike icon will not be named as a journalist, assigned professional credentials, or used to imply that a real person wrote the report. This constraint isolates appearance from false provenance claims. A short

comprehension check will distinguish "who presented the report" from "who drafted the report" before randomization.

3.2. Stimulus Development and Pretesting

The initial stimulus pool will cover 24 low-sensitivity factual topics, with one AI-drafted and one human-written report for each topic. Topics may include public transport, cultural events, science announcements, and community services. Each version will use the same verified fact sheet and a prespecified length range. Human writers will compose their version without generative drafting tools; AI versions will retain model and prompt records. Editing will be restricted to documented factual correction and formatting, with substantive rewriting triggering replacement of the stimulus. This defines AI-drafted material transparently without implying that editorial checking was absent.

Each participant will receive eight reports drawn without replacement from eight different topics, balanced between drafting sources. No participant will see both source versions of the same topic. Across participants, versions and topics will be balanced across the four experimental cells. The same text will appear under both presentation styles and both disclosure assignments across participants. Thus, treatment effects will not be confounded with differences between the actual texts shown in different interface conditions.

For $X = 1$, the label will identify the drafting source and editorial review using parallel wording: "Drafted with generative AI; fact-checked by an editor" or "Written by a human author; fact-checked by an editor." For $X = 0$, the matched tag will state "News report; fact-checked by an editor." The disclosure contrast therefore adds source information to a constant account of editorial checking. Fonts, placement, dimensions, display duration, and contrast will remain identical. The final Chinese wording will be fixed before recruitment.

An independent pretest will evaluate readability, factual equivalence, emotional intensity, icon humanlikeness, visual prominence, and the distinction between presenter and author. The pretest will not select items simply because they maximize the predicted recognition effect. Stimuli will be revised using prespecified quality criteria, then frozen. A manipulation can fail or introduce an unintended difference; either outcome calls for revision before the main experiment rather than post hoc reinterpretation of the construct.

3.3. Recruitment and Sample Size Planning

Recruitment will seek variation in age, education, and news use rather than relying exclusively on university students. Quotas can improve sample coverage but will not make an online panel a probability sample of the Chinese population. Compensation, recruitment channels, completion rates, and exclusions will be reported. No completed sample size is asserted in this protocol.

The final recruitment target will be set by simulation before data collection. The simulation will represent participants and reports as crossed sources of variation, include the planned number of reports per participant, and vary plausible interaction effects, mediator correlations, scale reliability, and attrition. Stimulus variation matters because adding participants cannot necessarily compensate for a narrow stimulus pool [18]. Power planning will prioritize the $X \times W$ contrast and the two moderated-mediation indices; rules based only on the number of questionnaire items will not be used. Published guidance on mediated-effect power further illustrates why an adequate sample for a main effect may be inadequate for a path product [19].

3.4. Measurement and Procedure

After consent, participants will complete eligibility checks and brief measures of prior generative-AI use and news use. Randomization will follow these baseline measures. In each report trial, the report and production tag will be removed before the authorship judgments and mediator items appear. The outcome is therefore immediate attribution after exposure to available cues, not unaided detection from an unlabeled text. During exposure, an identical publication-details control will reopen the same tag in both

disclosure conditions without adding new information. Clicks and dwell time will be recorded separately as inspection indicators; they will not be equated with visual attention or comprehension. To assess measurement reactivity, participants will also be randomly allocated to a fixed measurement order: mediator questions before authorship judgments or authorship judgments before mediator questions. Order will be included in the analysis, and treatment effects will be reported separately by order as a sensitivity check. Neither order establishes the causal direction between a measured mediator and the outcome.

Table 2 summarizes the measures. The Appendix provides draft English items for translation and cognitive testing. The scrutiny items are newly proposed operationalizations. Social-presence wording is informed by the construct literature but is adapted to the news interface; it will not be described as a previously validated scale in its present form. Similarly, a visual manipulation check is separate from the social-presence measure and from the source-recognition outcome. Measurement instruments developed for robots can inform item selection but require evaluation when transferred to a news interface [20].

Table 2. Construct Definitions and Measurement Boundaries.

Construct	Proposed measurement	Boundary to preserve
Source scrutiny M1	Three items on examining and considering evidence of drafting origin; behavioral inspection recorded separately.	Does not ask whether the report was AI-generated.
Social presence M2	Three items on the felt presence of a communicator and a socially engaging presentation.	Does not ask whether a human wrote the text.
AI attribution Y	Unmarked 0–100 slider, requiring an active response; separate binary classification.	Probability of AI drafting is not general trust or self-rated detection skill.
Presentation check	Humanlikeness of the interface icon, assessed after substantive outcomes.	Checks W without replacing randomized assignment with a perception score.
Secondary credibility	Separate appraisal of the report's credibility.	An evaluative outcome, excluded from the primary mediation model.

Credibility will be assessed separately, using wording informed by message-credibility measurement research [21]. It is included to test whether engagement, evaluation, and provenance recognition dissociate; it will not be inserted as an additional mediator after inspecting the results. Participants will receive a debrief explaining the production procedures and the distinction between factual accuracy and AI authorship.

3.5. Ethics and Transparency

Approval from the appropriate institutional ethics committee and informed consent will be obtained before recruitment. This prospective statement does not assert that approval or registration has already occurred. The study will use factual, low-sensitivity materials and disclose in consent that reports may have different production origins. Withholding an item-level authorship label will not involve displaying a false human byline. Identifiable panel information will be separated from research responses, and shared data will omit identifying fields.

The hypotheses, research questions, source definitions, stimuli, exclusions, sample-size simulation, outcome coding, and analysis code will be registered before the main study. Stimulus provenance and editing logs will accompany the materials, subject to licensing restrictions. These records are necessary to make an accuracy claim auditable and to prevent retrospective changes to what counts as AI-generated.

4. Planned Statistical Analysis

4.1. Data Quality and Measurement Evaluation

Exclusions will be limited to prespecified eligibility failures, duplicate participation, verified technical failures, and demonstrably invalid response patterns. Slow or inaccurate source recognition will not be an exclusion criterion. Failure to remember a disclosure will likewise not justify removal from the primary analysis: label uptake is part of the process under study. Participant flow and missingness will be reported by randomized condition. Primary mixed models will retain available trials under a conditional missing-at-random assumption. Because that assumption is unverifiable, sensitivity analyses will vary plausible missing outcomes and report whether condition-dependent dropout changes the substantive conclusions.

Before testing indirect associations, the study will examine item distributions, missing responses, internal consistency, and separation of the two mediator constructs. Confirmatory measurement models will compare the proposed two-factor structure with a one-factor alternative while accounting for repeated observations. Reliability estimates will be accompanied by uncertainty and will not substitute for construct validity [22]. Measurement comparability across presentation conditions will be assessed before interpreting group contrasts. If the mediators cannot be distinguished empirically, the analysis will report that limitation rather than relabeling overlapping measures as two validated mechanisms.

4.2. Total Effects and Recognition Performance

The randomized treatment analysis comes first. For AI-drafted reports, a mixed-effects model will estimate AI-attribution probability from X, W, and X × W, with prespecified baseline covariates and measurement order. Participant and topic variation will be modeled explicitly, with the two drafting versions nested within topic in analyses that include both sources. H2 will be evaluated in a separate mixed-effects model of social presence using the same treatment factors. Parallel analyses will examine human-written reports and binary classification. A combined model will include actual source and its interactions with the treatments, allowing improvement in recognition to be distinguished from a shift toward indiscriminate AI suspicion.

Classification reporting will include the proportion of AI reports correctly classified and the proportion of human reports incorrectly classified as AI. Overall probabilistic accuracy will be summarized using the Brier score, the mean squared difference between the reported probability, rescaled to 0–1, and the documented source indicator. This score reflects more than calibration alone. Calibration plots will separately compare stated AI probabilities with observed AI-source proportions using prespecified probability bins and participant-cluster uncertainty. Classification can remain unchanged while confidence becomes more extreme, so these outcomes will not be collapsed into one conclusion. An exploratory signal-detection summary may supplement these measures; extreme hit and false-alarm rates will require an explicit correction, and estimates based on few trials per person will be interpreted cautiously [23].

4.3. Moderated Parallel Mediation

The primary conditional process model concerns AI-drafted reports. Because X and W are assigned between participants, the mechanism of interest is a between-participant difference in average response to sampled AI reports. Mediator–outcome associations at that level must be separated from within-participant associations across reports. Treating all repeated ratings as independent observations would inflate precision and conflate these processes.

The explanatory mediation analysis will use participant means across their four AI-drafted reports. Each participant contributes one value for each mediator and for AI-attribution probability. The following linear equations include baseline covariates and measurement order in the vector C:

$$M_{1i} = i_1 + a_{11}X_i + a_{21}W_i + a_{31}X_iW_i + \gamma_1{}^{\mathrm{T}}C_i + e_{1i} \qquad (1)$$

$$M_{2i} = i_2 + a_{12}X_i + a_{22}W_i + a_{32}X_iW_i + \gamma_2{}^{\mathrm{T}}C_i + e_{2i} \quad (2)$$

$$Y_i = i_Y + c'X_i + dW_i + gX_iW_i + b_1M_{1i} + b_2M_{2i} + \delta^{\mathrm{T}}C_i + e_{Yi} \quad (3)$$

Here M1 and M2 are the participant means of source scrutiny and social presence, Y is the mean AI-attribution judgment, and X and W are coded 0 or 1. The mediator residuals may covary. The coefficients describe between-participant associations in average judgments over the assigned reports; they are not estimates of within-person psychological change. The explanatory mediation analysis will use participants with complete mediator and outcome records for all four AI trials. It therefore estimates associations among completers; condition-dependent completion or response-dependent missingness may introduce selection bias. Partial records will remain eligible for the primary randomized-effects analysis. Completion rates and a prespecified missing-data sensitivity analysis will be reported.

For mediator k, the conditional indirect association at presentation level w and the difference between presentation conditions are:

$$IE_k(w) = (a_{1k} + a_{3k}w)b_k \quad (4)$$

$$IMM_k = IE_k(1) - IE_k(0) = a_{3k}b_k \quad (5)$$

The two indices will be estimated in the same parallel model. This first-stage specification assumes that b1 and b2 do not vary with W; it does not include mediator-by-presentation interactions. A model relaxing that assumption would require a different expression for the conditional indirect association. A nonzero first-stage interaction alone does not establish moderated mediation, because the complete index also contains the relevant mediator–outcome coefficient. Likewise, an indirect association that is statistically distinguishable from zero in one group but not another does not establish a between-group difference. The contrast itself must be evaluated [17]. Parallel-path estimation also permits a direct contrast between the two indirect associations rather than ranking their separate significance levels [24].

Uncertainty will be estimated using 5,000 bootstrap resamples of participants, stratified by the eight treatment-and-measurement-order allocation combinations. Each resampled participant will retain the complete set of their trial records, and means and path products will be recomputed together. The same resample will be used for both mediators, preserving covariance when their indirect associations are contrasted. This procedure supports inference conditional on the stimulus pool; it does not account for uncertainty from sampling a new population of reports. The crossed participant–report analysis of randomized total effects remains the primary evidence about the interface contrast, and will not be described as validating the mediation mechanism.

The two moderated-mediation indices form one confirmatory family. Alongside ordinary 95% intervals, 97.5% intervals for each index will provide a Bonferroni-adjusted familywise assessment. The total interaction and conditional estimates will be reported with their magnitudes and uncertainty even when the indices are inconclusive. Unstandardized coefficients will be primary; standardized versions will be supplementary. For the 0–100 outcome, a path product is expressed in judgment-scale points under the fitted linear model, not as a risk difference in correct classification.

4.4. Robustness and Causal Limits

Robustness analyses will compare linear probability-judgment models with a fractional-response specification that accommodates endpoints, and compare their conclusions with a logistic model for binary classification. Simple linear path products will not be transferred to nonlinear models or interpreted as probability-scale causal indirect effects. Topic exclusions, measurement order, and prespecified baseline adjustment will be examined without selecting the specification that gives the smallest p value.

Randomization identifies effects of the assigned interface conditions under the usual experimental assumptions. It does not randomize scrutiny or social presence. Unmeasured post-exposure states, measurement reactivity, and reverse relationships can affect mediator–outcome associations. Causal mediation requires additional identification

assumptions, including adequate control of mediator–outcome confounding [25]. The measured path products will therefore be described as model-dependent indirect associations. The established limitations of cross-sectional mediation also rule out treating a single post-exposure measurement sequence as proof of temporal mechanism [26].

A planned report should permit several substantive outcomes: disclosure may improve recognition without either mediator accounting for the change; presentation may increase social presence without affecting recognition; or conditional indirect associations may point in opposing directions. A null interaction will be evaluated using its confidence interval and the prespecified smallest effect of interest, rather than being described automatically as evidence of no moderation. These possibilities are consequences of the model's distinctions, not failures to obtain the preferred narrative.

5. Discussion

5.1. Theoretical Contribution

The proposed framework connects MAIN's account of technological cues with an explicit test of production-origin recognition. It narrows the contribution to two independently varied interface features and two distinguishable psychological responses. The model does not infer detection ability from credibility ratings, and it does not interpret a humanlike social experience as an error by definition. Its principal theoretical value is specifying when those responses could diverge.

The treatment of anthropomorphism as a moderator makes the boundary condition explicit. Rather than asking only whether a humanlike presenter is liked, the model asks whether it changes the relationship between informative disclosure and a verifiable judgment. If the interaction is small with precise uncertainty, an interpretation would be that the tested disclosure remains effective across the two styles. If it is substantial, the mediator analysis can identify patterns consistent with differential scrutiny or social experience, while leaving causal mechanism claims appropriately qualified.

Including human-written reports also changes the meaning of success. A design that merely increases the frequency of AI judgments could appear effective in an all-AI stimulus set. The mixed-source benchmark exposes that possibility. The contribution is therefore about calibrated provenance judgments under interface variation, rather than maximizing suspicion toward automated news.

5.2. Implications for News Presentation

The framework suggests evaluating production transparency alongside readability and engagement. A disclosure should communicate which stage involved AI and should be tested for comprehension. Human authorship, AI drafting, presentation, and editorial checking are distinct roles. Combining them into an undifferentiated "AI-powered" label can leave readers unsure about what has actually been automated.

Practical recommendations will depend on the empirical results. If a humanlike presenter increases social presence while preserving accurate source recognition, social engagement and transparency may be compatible in that implementation. If recognition worsens, designers should test changes to label wording, prominence, or separation from presenter identity. The appropriate response is to evaluate the cue conflict directly; the present protocol does not justify a general ban on anthropomorphic interfaces or a claim that they reliably improve news understanding.

Accuracy findings would also need to be considered alongside false positives. Encouraging readers to classify human-written reports as AI-generated can misattribute labor and responsibility. A deployment decision should therefore examine both source categories, the uncertainty of the estimates, and calibration across different news topics. Credibility outcomes may inform that decision but cannot replace provenance performance.

5.3. Limitations and Extensions

This article presents a theoretical model and prospective protocol, not a completed experiment. The proposed operationalization examines a restrained visual contrast; it does not capture expressive speech, realistic video anchors, conversational memory, or adaptive dialogue. Those features could activate additional MAIN affordances and would require separate manipulations. The binary source benchmark is also deliberately limited. Hybrid drafting arrangements with substantial human revision should be studied as a distinct production category rather than forced into an ambiguous AI–human dichotomy.

An online panel and a bounded topic pool will limit generalizability. Repeated authorship judgments make provenance more salient than it may be during ordinary news use. The measurement-order comparison can diagnose some prompting effects but cannot reproduce every natural browsing context. Future studies could use unobtrusive exposure, delayed recognition, independently manipulated mediators, or field experiments to examine persistence and mechanisms. Such extensions should follow evidence about measurement quality and effect magnitude from the initial experiment.

6. Conclusion

Anthropomorphic presentation may alter how audiences use authorship disclosures, but a social response to an interface is not itself evidence of failed AI recognition. The proposed MAIN-based framework separates source scrutiny, social presence, and verifiable production-origin judgments within a moderated parallel mediation model. Its factorial protocol assigns disclosure and presentation independently, includes human-written comparison reports, and distinguishes judgment shifts from recognition accuracy. The resulting evidence could determine whether humanlike presentation weakens, strengthens, or leaves unchanged the effectiveness of truthful provenance information. Until that evidence is collected, the model specifies testable relationships rather than established effects.

Data Availability: No new participant data were collected or analyzed for this conceptual article and prospective protocol. The proposed hypotheses and procedures have not been reported as preregistered or empirically tested.

Appendix Draft Measurement Items: These English items specify the intended constructs. Chinese wording will undergo translation, independent review, and cognitive testing before the main study. All mediator items use seven response categories from strongly disagree to strongly agree. Item order will be randomized within each mediator block. Source scrutiny: "I examined information about how this report was produced." "I considered the evidence available about the report's drafting source." "I paid attention to details that could establish the origin of the report." These are proposed study-specific items; behavioral inspection will remain a separate measure. Social presence: "The presentation gave me a sense of a communicator being present." "The encounter felt socially engaging." "The interface created a sense of personal contact." The instructions will refer to the presentation experience rather than to the truth of the report or the humanity of its author. AI-attribution probability: "How likely is it that generative AI drafted the written text of this report?" Responses range from 0, definitely human-written, to 100, definitely AI-drafted. Participants must actively choose a value; the slider has no recorded default. Source classification: "Which production origin best describes the written text?" Response options are "AI-drafted" and "Human-written," with option order randomized. The probability question captures uncertainty that the forced classification does not. Presentation check: "How humanlike was the presenter's appearance?" Responses range from 1, not at all humanlike, to 7, extremely humanlike. Disclosure recall asks which production information was displayed and includes a "do not remember" option. Recall is a diagnostic measure, not an eligibility criterion.

References

1. C. Clerwall, "Enter the robot journalist: Users' perceptions of automated content," *Journalism Practice*, vol. 8, no. 5, pp. 519–531, 2014, doi: 10.1080/17512786.2014.883116.
2. A. Graefe, M. Haim, B. Haarmann, and H.-B. Brosius, "Readers' perception of computer-generated news: Credibility, expertise, and readability," *Journalism*, vol. 19, no. 5, pp. 595–610, 2018, doi: 10.1177/1464884916641269.
3. T. F. Waddell, "A robot wrote this? How perceived machine authorship affects news credibility," *Digital Journalism*, vol. 6, no. 2, pp. 236–255, 2018, doi: 10.1080/21670811.2017.1384319.

4. T. F. Waddell, "Attribution practices for the man-machine marriage: How perceived human intervention, automation metaphors, and byline location affect the perceived bias and credibility of purportedly automated content," *Journalism Practice*, vol. 13, no. 10, pp. 1255–1272, 2019, doi: 10.1080/17512786.2019.1585197.
5. S. S. Sundar, "The MAIN model: A heuristic approach to understanding technology effects on credibility," in *Digital Media, Youth, and Credibility*, M. J. Metzger and A. J. Flanagin, Eds. Cambridge, MA: MIT Press, 2008, pp. 73–100, doi: 10.1162/dmal.9780262562324.073.
6. S. S. Sundar, "Rise of machine agency: A framework for studying the psychology of human–AI interaction (HAII)," *Journal of Computer-Mediated Communication*, vol. 25, no. 1, pp. 74–88, 2020, doi: 10.1093/jcmc/zmz026.
7. M. Jakesch, J. T. Hancock, and M. Naaman, "Human heuristics for AI-generated language are flawed," *Proc. Natl. Acad. Sci.*, vol. 120, no. 11, Art. no. e2208839120, 2023, doi: 10.1073/pnas.2208839120.
8. J. Cohen, "Defining identification: A theoretical look at the identification of audiences with media characters," *Mass Communication and Society*, vol. 4, no. 3, pp. 245–264, 2001, doi: 10.1207/S15327825MCS0403_01.
9. M. K. Johnson, S. Hashtroudi, and D. S. Lindsay, "Source monitoring," *Psychological Bulletin*, vol. 114, no. 1, pp. 3–28, 1993, doi: 10.1037/0033-2909.114.1.3.
10. M. Haim and A. Graefe, "Automated news: Better than expected?" *Digital Journalism*, vol. 5, no. 8, pp. 1044–1059, 2017, doi: 10.1080/21670811.2017.1345643.
11. H. Jia, A. Appelman, M. Wu, and S. Bien-Aimé, "News bylines and perceived AI authorship: Effects on source and message credibility," *Comput. Hum. Behav.: Artif. Humans*, vol. 2, no. 2, Art. no. 100093, 2024, doi: 10.1016/j.chbah.2024.100093.
12. N. Epley, A. Waytz, and J. T. Cacioppo, "On seeing human: A three-factor theory of anthropomorphism," *Psychological Review*, vol. 114, no. 4, pp. 864–886, 2007, doi: 10.1037/0033-295X.114.4.864.
13. Y. Kim and S. S. Sundar, "Anthropomorphism of computers: Is it mindful or mindless?" *Comput. Hum. Behav.*, vol. 28, no. 1, pp. 241–250, 2012, doi: 10.1016/j.chb.2011.09.006.
14. W. Jang, J. W. Chun, S. Kim, and Y. W. Kang, "The effects of anthropomorphism on how people evaluate algorithm-written news," *Digital Journalism*, vol. 11, no. 1, pp. 103–124, 2023, doi: 10.1080/21670811.2021.1976064.
15. E. Konya-Baumbach, M. Biller, and S. von Janda, "Someone out there? A study on the social presence of anthropomorphized chatbots," *Comput. Hum. Behav.*, vol. 139, Art. no. 107513, 2023, doi: 10.1016/j.chb.2022.107513.
16. K. J. Preacher, D. D. Rucker, and A. F. Hayes, "Addressing moderated mediation hypotheses: Theory, methods, and prescriptions," *Multivariate Behav. Res.*, vol. 42, no. 1, pp. 185–227, 2007, doi: 10.1080/00273170701341316.
17. A. F. Hayes, "An index and test of linear moderated mediation," *Multivariate Behav. Res.*, vol. 50, no. 1, pp. 1–22, 2015, doi: 10.1080/00273171.2014.962683.
18. J. Westfall, D. A. Kenny, and C. M. Judd, "Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli," *J. Exp. Psychol.: Gen.*, vol. 143, no. 5, pp. 2020–2045, 2014, doi: 10.1037/xge0000014.
19. M. S. Fritz and D. P. MacKinnon, "Required sample size to detect the mediated effect," *Psychol. Sci.*, vol. 18, no. 3, pp. 233–239, 2007, doi: 10.1111/j.1467-9280.2007.01882.x.
20. C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *Int. J. Soc. Robotics*, vol. 1, no. 1, pp. 71–81, 2009, doi: 10.1007/s12369-008-0001-3.
21. A. Appelman and S. S. Sundar, "Measuring message credibility: Construction and validation of an exclusive scale," *J. Mass Commun. Q.*, vol. 93, no. 1, pp. 59–79, 2016, doi: 10.1177/1077699015606057.
22. J. K. Flake, J. Pek, and E. Hehman, "Construct validation in social and personality research: Current practice and recommendations," *Soc. Psychol. Pers. Sci.*, vol. 8, no. 4, pp. 370–378, 2017, doi: 10.1177/1948550617693063.
23. M. J. Hautus, "Corrections for extreme proportions and their biasing effects on estimated values of d′," *Behav. Res. Methods Instrum. Comput.*, vol. 27, no. 1, pp. 46–51, 1995, doi: 10.3758/BF03203619.
24. K. J. Preacher and A. F. Hayes, "Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models," *Behav. Res. Methods*, vol. 40, no. 3, pp. 879–891, 2008, doi: 10.3758/BRM.40.3.879.
25. K. Imai, L. Keele, and D. Tingley, "A general approach to causal mediation analysis," *Psychol. Methods*, vol. 15, no. 4, pp. 309–334, 2010, doi: 10.1037/a0020761.
26. S. E. Maxwell and D. A. Cole, "Bias in cross-sectional analyses of longitudinal mediation," *Psychol. Methods*, vol. 12, no. 1, pp. 23–44, 2007, doi: 10.1037/1082-989X.12.1.23.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of Publisher and/or the editor(s). Publisher and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.