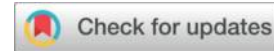




Algorithmically Mediated Authenticity: A Conceptual Framework for Digital Identity, Reputation, and Trust in the Era of Generative AI



Fatima Arab

Faculty of Social Sciences and Humanities, Communications Department, Akli Mohand Oulhadj University, Bouira (Algeria)

f.arab@univ-bouira.dz

Seba Fadila

Faculty of Social Sciences and Humanities, Communications Department, Akli Mohand Oulhadj University, Bouira (Algeria)

f.seba@univ-bouira.dz

Nouari Aicha

Faculty of Social Sciences and Humanities, Communications Department, Akli Mohand Oulhadj University, Bouira (Algeria)

a.nouari@univ-bouira.dz

Received: 25-03-2026

Accepted: 22-09-2026

Published: 08-10-2026

Abstract:

Generative AI is reshaping how digital identities are produced, displayed, and evaluated. Online self-presentation can be created or optimized by AI, while recommendation systems govern visibility. Research on self-presentation, algorithmic visibility, human-machine communication, authenticity, and source credibility remains fragmented. This conceptual article develops the Algorithmically Mediated Authenticity (AMA) model to explain how perceived AI assistance, perceived algorithmic mediation of visibility, and perceived strategic self-presentation influence perceived authenticity, online reputation, and interpersonal trust. It identifies disclosure quality, platform transparency, and audience AI literacy as boundary conditions. Based on an integrative review, the article formulates ten research propositions and proposes a survey-experimental validation strategy with a preliminary item pool. The model distinguishes content authorship from algorithmic visibility and offers implications for platform governance, AI disclosure, reputation management, and digital literacy.

Keywords; Digital identity; Generative AI; Algorithmic mediation; Authenticity; Digital trust.

1. Introduction

Online self-presentation is increasingly produced with the participation of artificial intelligence. Generative systems built on large language models, including those underlying ChatGPT, Gemini, Copilot, and Claude, together with image and voice generation tools, can draft a dating profile, polish a professional biography, generate a portrait, or compose an entire stream of social media content on a person's behalf (Hancock et al., 2020, pp. 89–92). This assistance is frequently

invisible to the audience: large-sample experimental evidence shows that people cannot reliably distinguish AI-generated self-presentations from human-written ones, and that their attempts to do so rest on intuitive but flawed heuristics that AI systems can exploit (Jakesch et al., 2023, pp. 1–2, 6–8). AI-mediated communication, understood as interpersonal communication that is not merely transmitted by technology but modified, augmented, or generated by it, has thus moved from a niche design question to an ordinary condition of digital life (Hancock et al., 2020, pp. 89–92).

The phenomenon matters both scientifically and socially. Scientifically, it strikes at an assumption embedded in decades of self-presentation research: that the identity performance under evaluation was authored by the human it represents. Socially, it matters because authenticity judgments carry consequences: they shape whom audiences trust, hire, book, follow, and believe (Jakesch et al., 2019, pp. 1–2, 8–10; Ohanian, 1990, pp. 41–46). Synthetic media affects those judgments even without deceiving anyone; experimental evidence shows that exposure to deepfakes primarily breeds uncertainty about what is real, and that this uncertainty in turn erodes trust in mediated content (Vaccari & Chadwick, 2020, pp. 1–2, 8–10). Regulators have begun responding: the European Union's Artificial Intelligence Act imposes transparency obligations on providers of AI systems that generate synthetic content (European Union, 2024, Art. 50), and the Coalition for Content Provenance and Authenticity (C2PA) is developing technical standards for signaling how digital content was produced (C2PA, n.d., "Technical Specifications" section). These governance developments indicate that the question of who, or what, authored an online identity performance is now treated as consequential by institutions as well as by audiences.

Previous research offers substantial but fragmented resources for understanding this condition. Self-presentation scholarship explains how individuals strategically construct identity for imagined audiences under conditions of context collapse (Goffman, 1959, pp. 17–25; Marwick & boyd, 2011, p. 114). Research on algorithmic power explains how platform ranking systems govern which identity performances become visible at all (Bucher, 2012, p. 1164; Gillespie, 2014), and a growing empirical literature documents wide variation in users' awareness of that algorithmic mediation (Gran et al., 2021, pp. 1781–1784, 1790–1792; Hargittai et al., 2020, pp. 765–769). Human-machine communication scholarship establishes that AI systems increasingly occupy the position of communicator rather than channel (Guzman & Lewis, 2020, pp. 72–76). And an emerging experimental literature on AI-mediated communication demonstrates that perceived AI involvement measurably affects trust (Jakesch et al., 2019, pp. 1–2, 8–10) while remaining largely undetectable in practice (Jakesch et al., 2023, pp. 1–2, 6–8; Köbis & Mossink, 2021, pp. 1–2, 7–9).

What remains missing is integration. Our review, described in Section 2, did not identify an existing framework that simultaneously integrates the perceived authorship of identity content, the algorithmic mediation of its visibility, and the authenticity judgments through which audiences convert both into reputation and trust. Each tradition brackets precisely the variable the others foreground.

To address this gap, this article develops the Algorithmically Mediated Authenticity (AMA) model, which integrates self-presentation theory, algorithmic visibility research, human-machine communication, and source credibility scholarship. The model explains how perceived AI involvement and algorithmic mediation shape authenticity, reputation, and trust, while accounting for content-level disclosure quality, platform transparency, and audience AI literacy as distinct boundary conditions. Because the model has not yet been empirically tested, the article is explicitly conceptual: it formulates research propositions rather than hypotheses, presents a preliminary item pool rather than a validated instrument, and closes by specifying the empirical strategy through which the propositions could be validated in future research.

2. Conceptual Development and Literature Review Approach

This study adopts an integrative conceptual review approach rather than a systematic review protocol; no claim of exhaustive coverage is made, and the review's purpose was theory construction rather than evidence synthesis. Literature was identified through structured searches of Scopus, Web of Science, the ACM Digital Library, ScienceDirect, and the SAGE and Taylor & Francis journal collections, supplemented by backward and forward citation chaining from foundational works (Google Scholar was used for chaining). Search terms combined the focal constructs in varying pairs: "digital identity," "online self-presentation," "authenticity," "generative AI," "AI-mediated communication," "human-machine communication," "algorithm awareness," "algorithmic visibility," "source credibility," "trust," and "AI disclosure."

The period covered begins with foundational theoretical works on self-presentation (Goffman, 1959, pp. 17–25) and extends through empirical scholarship published up to the final database search, conducted in July 2026, with particular attention to the 2019–2026 window in which experimental research on AI-mediated self-presentation and AI-content labeling emerged. Inclusion criteria were: peer-reviewed journal articles, full papers in ranked conference proceedings (e.g., CHI), and scholarly monographs from university or established academic presses; and substantive theoretical or empirical treatment of at least one focal construct (identity performance, algorithmic visibility, machine communicative agency, authenticity, credibility, or trust). Excluded were non-peer-reviewed commentary, purely technical detection-algorithm papers lacking an audience or communication dimension, and sources whose bibliographic details could not be verified.

Synthesis was narrative and integrative: rather than organizing the literature chronologically or by discipline, we organized it by the communicative function each tradition explains, namely the production of identity content (self-presentation and AI-mediated communication), the visibility of that content (algorithmic power and algorithm awareness), and its evaluation by audiences (source credibility and trust). The four theoretical traditions foregrounded in Section 3 were selected because each supplies a construct the others lack, and because their intersection is precisely where generative AI operates. On the basis of this review, bounded as described, the article claims a gap: our review did not

identify a framework that simultaneously integrates these three functions; the claim should be read as relative to this search strategy rather than as absolute.

3. Theoretical Background

3.1 Self-Presentation and the Problem of Authorship

The dramaturgical account of self-presentation treats identity as a continual performance calibrated to a given audience, and its extension to networked environments has been one of communication science's most productive moves (Goffman, 1959, pp. 17–25). Research on Twitter users facing "context collapse," or the flattening of multiple audiences into one, showed that users construct an imagined audience and adjust disclosure, humor, and self-branding accordingly, often using techniques resembling the strategic self-commodification of micro-celebrity (Marwick & boyd, 2011, p. 114). Notably, several respondents treated the very act of consciously addressing an audience as inauthentic, insisting that they presented themselves for their own sake rather than for a calculated public (Marwick & boyd, 2011, p. 119); authenticity, in this tradition, is a socially constructed judgment rather than an objective property. This argument was later extended to networked publics generally, showing that persistence, searchability, and replicability make context collapse structural rather than incidental (boyd, 2014, pp. 8–11).

Self-presentation theory thus explains how individuals strategically construct identity for an imagined audience, but it does not account for uncertainty surrounding machine participation in the production of that identity. The theory's implicit model of authorship is singular and human: a person decides what to reveal, conceal, or embellish. Generative AI reopens authorship as a variable at exactly the point where the theory had closed it; the audience must now judge not only whether a performance is strategically curated, but whether it was humanly produced at all.

3.2 Algorithmic Visibility, Algorithm Awareness, and the Governance of Attention

Where self-presentation theory examines what is performed, algorithmic visibility research examines what is seen. An analysis of Facebook's EdgeRank algorithm introduced the "threat of invisibility": users perform not only for human audiences but for an algorithmic gatekeeper whose criteria are opaque, and the governing anxiety is not being watched but constantly disappearing (Bucher, 2012, pp. 1164, 1174). Users respond by learning to play the game, strategically producing the interactions the algorithm rewards (Bucher, 2012, p. 1174), and by developing an "algorithmic imaginary," or folk theories of algorithmic operation that shape behavior regardless of accuracy (Bucher, 2017, pp. 30–35, 40–42). Ranking systems are therefore not neutral conduits but active participants in defining what counts as socially significant (Gillespie, 2014, pp. 167–172). Subsequent empirical work has shown that awareness of this algorithmic mediation is unevenly distributed across populations, to the point of constituting a potential new dimension of the digital divide (Gran et al., 2021, pp. 1781–1784, 1790–1792); that measuring people's algorithm skills poses distinct methodological challenges (Hargittai et al., 2020, pp. 765–769); and that

awareness can itself be deliberately heightened through interface interventions (Fouquaert & Mechant, 2022, pp. 1773–1778, 1783–1785). These findings motivate treating perceived, rather than actual, algorithmic mediation as the audience-side construct.

This tradition establishes that visibility is co-produced by infrastructures with their own logics, yet it treats the ranked content as an already-authored artifact. It theorizes the second stage of a pipeline, how content becomes visible after it has been produced, while taking the first stage, production itself, as settled. The AMA model connects the two stages: audiences who believe algorithms shape what they see may also grow suspicious about how what they see was made, and both beliefs plausibly feed the same authenticity judgment.

3.3 Human–Machine Communication and AI-Mediated Communication

Human–machine communication (HMC) scholarship supplies the missing theory of machine authorship. AI technologies no longer fit the classical assumption that technology merely mediates human-to-human exchange; conversational agents and language-generation systems increasingly occupy the structural position of communicator (Guzman & Lewis, 2020, pp. 72–76). Research on mobile virtual assistants has also shown that users apply source-orientation heuristics to machines that closely parallel those documented for human sources (Guzman, 2019, pp. 344–348). Yet HMC has predominantly examined direct interactions between users and AI systems, specifically the dyad of a person talking to a machine. The situation at issue here is different: an AI system contributes to a human identity performance that is then presented, often without disclosure, to a third-party audience.

That triadic situation is the province of AI-mediated communication (AI-MC), defined as interpersonal communication in which an intelligent agent operates on behalf of a communicator by modifying, augmenting, or generating messages (Hancock et al., 2020, pp. 89–92). Two experimental findings from this literature anchor the present framework. First, three experiments with 527 participants evaluating Airbnb host profiles found that hosts believed to have used AI-written profiles were rated less trustworthy, but only when participants believed they were viewing a mixed set of AI- and human-written profiles, a conditional pattern termed the Replicant Effect (Jakesch et al., 2019, pp. 1–2, 8–10). Second, six experiments with 4,600 participants found that people could not reliably detect AI-generated self-presentations in professional, hospitality, and dating contexts; judgments rested on flawed heuristics, such as associating first-person pronouns or family topics with human authorship, that AI systems can exploit to appear "more human than human" (Jakesch et al., 2023, pp. 1–2, 6–8). Converging evidence extends beyond self-presentation: participants could not reliably distinguish AI-generated from human-written poetry (Köbis & Mossink, 2021, pp. 1–2, 7–9), and similar evaluation failures have been documented across generated text genres (Clark et al., 2021, pp. 7282–7285, 7290–7293).

Together these findings establish two premises on which the AMA model rests: perceived AI involvement affects trust conditionally rather than uniformly, and perception, not detection, is the operative variable, since actual detection is

unreliable. The present framework connects the self-presentation and HMC/AI-MC traditions by examining situations in which AI contributes to a human identity performance presented to a third-party audience, and by asking how the audience's beliefs about that contribution restructure authenticity, reputation, and trust.

3.4 Source Credibility and Authenticity Judgments

The evaluative side of the model draws on credibility and authenticity research. Authenticity is not an objective property of an entity but an attribution made by an audience relative to a referent, or a judgment about whether an entity is true to itself or to its claimed category (Lehman et al., 2019, pp. 3–8); this attributional understanding underwrites the model's decision to define authenticity as perceived. Source credibility research identifies expertise, trustworthiness, and attractiveness as core dimensions of communicator evaluation (Ohanian, 1990, pp. 41–46). The MAIN model is particularly relevant because its agency heuristic concerns the audience's attribution of who or what is responsible for content; online credibility judgments are shaped by cues about modality, agency, interactivity, and navigability rather than by message content alone (Sundar, 2008, pp. 73–81, 84–86). The agency heuristic thereby anticipates the authorship question that generative AI now poses at the level of infrastructure. It also implies that audiences differ in how they apply authorship cues, which connects credibility research to AI literacy, understood as competencies for understanding, interacting with, and critically evaluating AI systems (Long & Magerko, 2020, pp. 2–7; Ng et al., 2021, pp. 2–8). What credibility research has not yet done is embed the agency heuristic in a model that also includes algorithmic visibility beliefs and self-presentation strategy; this is the integration the AMA model undertakes.

3.5 Online Reputation as Accumulated Social Evaluation

Online reputation differs from an immediate credibility judgment because it aggregates visible evaluations from prior audiences. Online feedback mechanisms convert ratings, comments, recommendations, and interaction histories into public signals that help users form expectations and trust in otherwise unfamiliar actors (Dellarocas, 2003, pp. 1407–1411). In the AMA model, Perceived Online Reputation therefore captures socially accumulated standing inferred from visible social-evaluation cues, rather than merely the focal evaluator's personal impression. This distinction supports the proposed POR→IT path because reputation cues summarize prior evaluations that audiences use when deciding whether to rely on a digital identity.

4. Research Gap

Three specific gaps follow from the preceding review. First, a construct gap: our review did not identify a widely used measurement framework that operationalizes perceived AI assistance in self-presentation as a variable distinct from technology-acceptance constructs, which measure a user's willingness to adopt a tool, and from source-credibility constructs, which measure judgments of an already-identified source. Second, a mechanism gap: the AI-assistance–trust relationship is conditional on comparative context, but the relevant boundary conditions remain insufficiently specified (Jakesch et al., 2019, pp. 1–2, 8–10).

Content-level disclosure quality, platform-level transparency, and audience AI literacy are plausible but theoretically distinct conditions suggested by disclosure research, the MAIN model's agency heuristic, and AI literacy scholarship (Altay & Gilardi, 2024, pp. 1–3, 7–8; Lin & Zhang, 2026, pp. 1–3, 14–17; Long & Magerko, 2020, pp. 2–7; Ng et al., 2021, pp. 2–8; Schilke & Reimann, 2025, pp. 1–2, 12–14; Shi & Jiang, 2026, pp. 1, 5, 14–16; Sundar, 2008, pp. 73–81, 84–86), yet they have not been integrated into one model of digital identity evaluation. Third, an integration gap: self-presentation, algorithmic visibility, and human-machine communication scholarship have not, to the best of our review, been combined into a single structure, despite clear conceptual adjacency; all three concern who controls what becomes visible and how it is judged. Generative AI operates precisely at the seams between these traditions: in the authorship of self-presentation, in the algorithmic ranking of the authored content, and in the ambiguous communicative status of the AI system itself. A framework that treats these seams as its central object of study is what communication and information science currently lack.

5. The Algorithmically Mediated Authenticity Model

The AMA model proposes three independent constructs, Perceived AI Assistance (PAA), Perceived Algorithmic Mediation of Visibility (PAMV), and Perceived Strategic Self-Presentation (PSS), as antecedents of a central mediating construct, Perceived Authenticity (PAU). Perceived authenticity, in turn, is proposed to shape two related outcomes, Perceived Online Reputation (POR) and Interpersonal Trust (IT), with reputation additionally feeding trust. Three boundary conditions are specified around the PAA→PAU relationship. Content-Level Disclosure Quality (CDQ) captures the specificity and explanatory detail of the label attached to a particular item. Platform Transparency (PT) captures the broader consistency, visibility, and interpretability of AI-disclosure practices within the platform environment in which the stimulus is presented. Audience AI Literacy (AAL) captures user-side competencies and is modeled as a higher-order composite formed by self-perceived literacy and objective AI knowledge. Figure 1 depicts the model; Table 1 defines each construct and its theoretical anchor.

"Perceived Algorithmic Mediation of Visibility" is preferred over the shorter "algorithmic presence" to avoid confusion with social presence and telepresence constructs, and because the construct's definition and items concern specifically the role of ranking and amplification systems in visibility. "Perceived Strategic Self-Presentation" is specified as a unipolar construct, meaning the perceived degree of strategic, branded, calculated presentation, rather than a bipolar authentic-versus-strategic continuum, because the proposed items measure only the strategic pole. Content-Level Disclosure Quality is distinguished from Platform Transparency because a label attached to one item is not equivalent to the platform's broader disclosure environment. Audience AI Literacy is treated as a higher-order construct because perceived familiarity and objectively demonstrated knowledge may contribute differently to audience judgment. In Figure 1, P10 is represented as the conditional interaction of Audience AI Literacy with Content-Level Disclosure Quality on the PAA→PAU path.

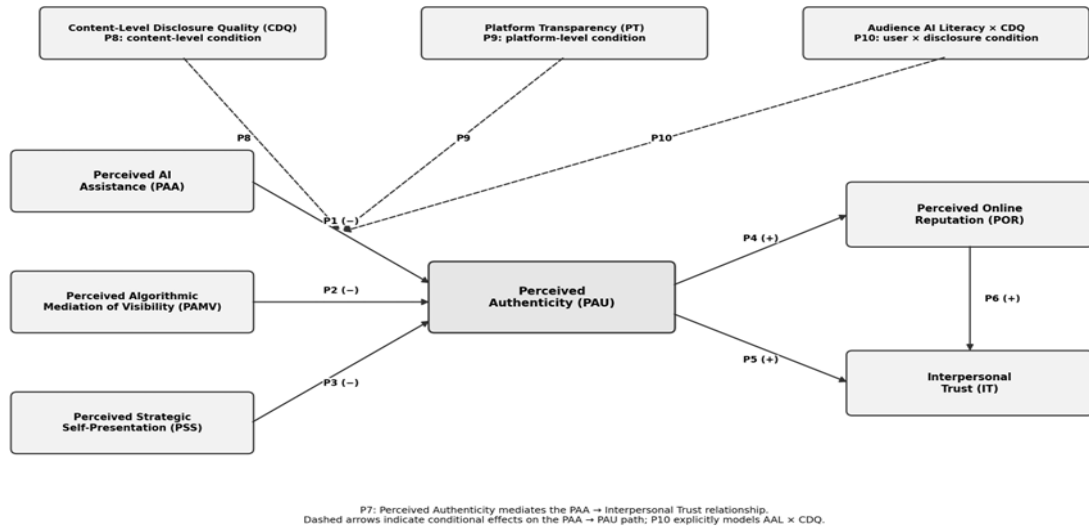


Figure 1. The Algorithmically Mediated Authenticity (AMA) conceptual model. Dashed arrows indicate conditional effects on the PAA→PAU path; P10 represents the AAL×CDQ interaction.

Table 1. Construct definitions and theoretical anchors.

Construct	Definition	Theoretical anchor
Perceived AI Assistance (PAA)	The degree to which an audience believes generative AI tools contributed to the content through which an identity is presented.	AI-mediated communication (Hancock et al., 2020, pp. 89–92; Jakesch et al., 2019, pp. 1–2, 8–10; Jakesch et al., 2023, pp. 1–2, 6–8)
Perceived Algorithmic Mediation of Visibility (PAMV)	The degree to which an audience believes algorithmic ranking, recommendation, and amplification systems shape the visibility of an identity, independent of its authored content.	Algorithmic power and visibility (Bucher, 2012, p. 1164; Bucher, 2017, pp. 30–35, 40–42; Gillespie, 2014, pp. 167–172); algorithm awareness (Gran et al., 2021, pp. 1781–1784, 1790–1792; Hargittai et al., 2020, pp. 765–769)
Perceived Strategic Self-Presentation (PSS)	The degree to which an audience perceives an identity performance as strategic, branded, and calculated for effect.	Self-presentation theory (Goffman, 1959, pp. 17–25); context collapse and micro-celebrity (Marwick & boyd, 2011, p. 114)
Perceived Authenticity (PAU)	Audience judgment that presented identity content genuinely reflects the represented person's own agency and voice.	Authenticity as audience attribution (Lehman et al., 2019, pp. 3–8); source credibility (Ohanian, 1990, pp. 41–46); MAIN model agency heuristic (Sundar, 2008, pp. 73–81, 84–86)

Construct	Definition	Theoretical anchor
Perceived Online Reputation (POR)	Audience assessment of an identity's socially accumulated standing, inferred from visible social evaluation cues (ratings, comments, followers, recommendations).	Online feedback and reputation mechanisms (Dellarocas, 2003, pp. 1407–1411); source credibility theory (Ohanian, 1990, pp. 41–46)
Interpersonal Trust (IT)	Audience willingness to rely on the presented identity, expressed as confidence in the information provided, willingness to accept vulnerability, and perceived reliability.	Perceived trustworthiness in AI-MC (Jakesch et al., 2019, pp. 1–2, 8–10)
Content-Level Disclosure Quality (CDQ) (moderator)	The clarity, specificity, and explanatory detail of the AI label attached to a particular item, including whether it distinguishes limited AI assistance from full AI generation and indicates human oversight.	AI-disclosure and labeling research (Altay & Gilardi, 2024, pp. 1–3, 7–8; Lin & Zhang, 2026, pp. 1–3, 14–17; Schilke & Reimann, 2025, pp. 1–2, 12–14; Shi & Jiang, 2026, pp. 1, 5, 14–16)
Platform Transparency (PT) (moderator)	The audience-perceived AI-disclosure environment of the platform presented in the study: the visibility, understandability, and consistency of provenance and AI-labeling practices across content.	Platform governance and content provenance (C2PA, n.d., "Technical Specifications" section); EU AI Act transparency obligations (European Union, 2024, Art. 50)
Audience AI Literacy (AAL) (moderator)	A higher-order user competency construct formed by Self-Perceived AI Literacy and Objective AI Knowledge.	AI literacy competency frameworks (Long & Magerko, 2020, pp. 2–7; Ng et al., 2021, pp. 2–8); MAIN model agency heuristic (Sundar, 2008, pp. 73–81, 84–86)

The model deliberately excludes constructs whose addition would outpace its measurement base. Source credibility appears as the theoretical anchor of perceived authenticity rather than as a separate mediating construct, and no direct path from perceived algorithmic mediation to trust is proposed; both extensions are identified in Section 10 as directions for future elaboration once the central structure has been empirically stabilized. CDQ is modeled as a content-level experimental condition, PT as a platform-level perceived environment, and AAL as a higher-order user competency construct, thereby avoiding the conflation of distinct levels of analysis.

6. Research Propositions

Because the AMA model has not yet been empirically tested, the relationships it specifies are formulated as research propositions rather than hypotheses. Each proposition is developed from the theoretical and empirical material reviewed

above; upon data collection through the strategy described in Section 7, these propositions can be converted into formal, testable hypotheses.

6.1 Perceived AI Assistance and Authenticity

The Replicant Effect demonstrates that audiences penalize communicators believed to have used AI in self-presentation, and that this penalty is conditional on the evaluative context in which the belief arises (Jakesch et al., 2019, pp. 1–2, 8–10). At the same time, detection of AI involvement is unreliable (Jakesch et al., 2023, pp. 1–2, 6–8; Köbis & Mossink, 2021, pp. 1–2, 7–9), which means the operative variable is the audience's perception of AI assistance rather than its actual presence. When that perception forms in the absence of any disclosure, the audience is positioned to interpret AI involvement as concealment, which credibility theory associates with reduced trustworthiness attributions (Ohanian, 1990, pp. 41–46).

Proposition 1 (P1). Perceived AI assistance in online self-presentation is negatively associated with perceived authenticity.

6.2 Perceived Algorithmic Mediation and Authenticity

The threat-of-invisibility account implies that audiences understand visibility as something achieved through strategic conformity to algorithmic logic rather than earned through unmediated expression (Bucher, 2012, p. 1174), and algorithm-awareness research confirms that many users hold explicit beliefs about such mediation, though unevenly across populations (Gran et al., 2021, pp. 1781–1784, 1790–1792). An identity whose prominence the audience attributes to algorithmic amplification is, by that attribution, an identity whose presence has been engineered; engineering sits uneasily with the folk concept of authenticity, in which calculated audience-seeking is itself read as inauthentic (Marwick & boyd, 2011, p. 119).

Proposition 2 (P2). Higher perceived algorithmic mediation of an identity's visibility is associated with lower perceived authenticity.

6.3 Perceived Strategic Self-Presentation and Authenticity

A folk opposition between strategic self-branding and authenticity has been documented: audiences and users alike read overtly promotional, branded self-presentation as calculated, and calculated presentation as less authentic (Marwick & boyd, 2011, pp. 114, 119). Perceived strategic self-presentation should therefore function as a direct antecedent of authenticity judgments.

Proposition 3 (P3). Higher perceived strategic self-presentation is associated with lower perceived authenticity.

6.4 Authenticity, Reputation, and Trust

Source credibility research establishes trustworthiness as a core dimension of communicator evaluation (Ohanian, 1990, pp. 41–46), and the AI-MC evidence shows that authenticity-relevant beliefs translate directly into trust outcomes (Jakesch et al., 2019, pp. 1–2, 8–10). Within the model, perceived authenticity is the evaluative hinge: it feeds the audience's assessment of the identity's socially accumulated standing (reputation) and its willingness to rely on the identity (trust), while reputation, as a more stable, socially shared evaluation inferred from visible cues such as ratings, comments, and follower counts, additionally informs

trust. Because perceived AI assistance is proposed to lower authenticity (P1) and authenticity is proposed to raise trust (P5), the model implies that authenticity carries the effect of AI-assistance perceptions through to trust.

Proposition 4 (P4). Perceived authenticity positively influences perceived online reputation.

Proposition 5 (P5). Perceived authenticity positively influences interpersonal trust.

Proposition 6 (P6). Perceived online reputation positively influences interpersonal trust.

Proposition 7 (P7). Perceived authenticity mediates the relationship between perceived AI assistance and interpersonal trust.

6.5 Boundary Conditions: Disclosure Quality, Platform Transparency, and AI Literacy

The conditionality of the Replicant Effect indicates that audience responses to AI assistance depend on contextual cues about its prevalence and legitimacy (Jakesch et al., 2019, pp. 1–2, 8–10). Three boundary conditions should be distinguished. First, content-level disclosure quality refers to the specificity and explanatory detail of the label attached to a particular item. Evidence shows that disclosure can produce skepticism or authenticity penalties when it implies full automation, while its effects vary with label design, information veracity, audience attitudes, and perceived authenticity (Altay & Gilardi, 2024, pp. 1–3, 7–8; Lin & Zhang, 2026, pp. 1–3, 14–17; Schilke & Reimann, 2025, pp. 1–2, 12–14; Shi & Jiang, 2026, pp. 1, 5, 14–16). Second, platform transparency concerns the broader environment in which disclosure rules are consistently visible, understandable, and applied across content (C2PA, n.d., "Technical Specifications" section; European Union, 2024, Art. 50). Third, audience AI literacy may improve calibration when AI use is openly disclosed and proportionate, but greater literacy may also make concealed or misleading involvement more salient. Its moderating effect should therefore be tested conditionally across disclosure-quality settings (Jakesch et al., 2023, pp. 1–2, 6–8; Long & Magerko, 2020, pp. 2–7; Ng et al., 2021, pp. 2–8; Sundar, 2008, pp. 73–81, 84–86).

Proposition 8 (P8). Content-level disclosure quality moderates the relationship between perceived AI assistance and perceived authenticity. A clear, contextualized label that distinguishes limited AI assistance from full AI generation and explains human oversight is expected to attenuate authenticity loss relative to a generic AI-generated label.

Proposition 9 (P9). Platform transparency moderates the relationship between perceived AI assistance and perceived authenticity. A clear, visible, and consistently applied platform disclosure environment is expected to improve audience calibration, whereas inconsistent or poorly explained transparency practices may intensify authenticity loss.

Proposition 10 (P10). Audience AI literacy moderates the relationship between perceived AI assistance and perceived authenticity, and this moderation varies by content-level disclosure quality. Higher literacy is expected to attenuate authenticity loss under clear and proportionate

disclosure, but may intensify negative judgments when AI involvement is concealed or misleading.

7. Proposed Empirical Validation Strategy

The AMA model has not yet been empirically tested. This section proposes a research design through which the model may be validated in future research. The design would follow a post-positivist paradigm and a deductive approach: an explanatory, cross-sectional survey supplemented by an embedded 4×2 factorial experiment manipulating four content-level authorship-label conditions and two platform-transparency environments, in the spirit of the vignette logic used to isolate the Replicant Effect (Jakesch et al., 2019, pp. 1–2, 8–10). Causal inference would apply to the experimentally manipulated label and platform-transparency conditions; the remaining structural relationships would be tested correlationally and interpreted accordingly.

7.1 Population and Sampling

The target population would consist of active social media users aged 18 and above who maintain at least one public-facing profile on a mainstream platform. The study would employ quota-based purposive sampling across predefined participant groups, including university students, early-career professionals, incubator-affiliated entrepreneurs, and a general-population online panel, with quotas on age and platform use; this composition would situate the first validation in a North African higher-education and entrepreneurship context, a setting underrepresented in the AI-MC literature, while the general-panel stratum would permit comparison with less digitally intensive users. Snowball referral within quota cells could supplement recruitment, with the attendant self-selection limitations acknowledged in reporting.

Sample size would be determined through a priori power analysis and Monte Carlo simulation anchored to the smallest substantively meaningful effect in the model, particularly the interaction terms (P8–P10), together with the mediation pathway (P7), the eight cells of the 4×2 factorial experiment, the split-sample EFA/CFA sequence, and an anticipated 15–20% rate of exclusions and incomplete responses. Simulation-based estimates for moderated-mediation structures of this size typically indicate required samples well above common heuristics; the ten-times rule, being only a rough heuristic that may underestimate requirements, would not serve as the primary justification. A preliminary recruitment target of 800–1,000 completed responses before exclusions is proposed, with the final minimum determined through Monte Carlo simulation.

7.2 Instrument, Pilot Study, and Measurement Quality

All constructs would be measured with multi-item, 7-point Likert scales (the preliminary item pool appears in the Appendix). Some items would be adapted from established measures: perceived authenticity, reputation, and trust items draw on credibility dimensions, online feedback research, and a trustworthiness measure (Dellarocas, 2003, pp. 1407–1411; Jakesch et al., 2019, pp. 1–2, 8–10; Ohanian, 1990, pp. 41–46). Items for perceived algorithmic mediation and audience AI literacy are newly developed from the literature and would undergo systematic content, construct, convergent, discriminant, and criterion validation

before substantive use (Gran et al., 2021, pp. 1781–1784, 1790–1792; Hargittai et al., 2020, pp. 765–769; Long & Magerko, 2020, pp. 2–7; Ng et al., 2021, pp. 2–8). The CDQ and PT items would primarily function as manipulation checks, verifying that participants perceived the contextualized label as clearer than the generic label and the high-transparency platform environment as clearer, more visible, and more consistent than the low-transparency environment. A pilot study with 30–40 respondents would precede full deployment to assess item clarity, completion time, manipulation strength, and preliminary reliability; items would be revised or dropped if corrected item-total correlations fell below .40 or cognitive-interview feedback indicated ambiguity. Content validity would be established through expert-panel review by at least five communication or information-science scholars.

Audience AI Literacy would be modeled as a higher-order composite formed by two lower-order dimensions: Self-Perceived AI Literacy (SPAIL), measured through confidence and familiarity items, and Objective AI Knowledge (OAIK), measured through accuracy-scored knowledge questions. The two dimensions would be scored and validated separately because perceived familiarity and demonstrated knowledge need not covary. If collinearity and indicator-weight diagnostics support aggregation, AAL would be estimated as a formative higher-order composite; otherwise, SPAIL and OAIK would be retained as separate moderators. This specification is consistent with competency-based accounts of AI literacy (Long & Magerko, 2020, pp. 2–7; Ng et al., 2021, pp. 2–8). Because several measures are newly developed, a split-sample validation sequence is proposed: exploratory factor analysis on a first sample to establish dimensionality, confirmatory factor analysis on a second, independent sample to confirm the measurement structure, and only then structural modeling to test the theoretical relationships. At the confirmatory stage, convergent and discriminant validity would be assessed through average variance extracted, the Fornell–Larcker criterion, and the heterotrait–monotrait ratio (Fornell & Larcker, 1981, pp. 45–46; Henseler et al., 2015, pp. 120–127); internal consistency would be assessed through Cronbach's alpha and composite reliability against the conventional .70 threshold (Nunnally & Bernstein, 1994, p. 265). Discriminant validity between perceived authenticity, trust, and reputation would receive particular scrutiny given their conceptual proximity; the item pool has been drafted to separate them (authenticity as genuineness of expression; trust as reliance, confidence, and willingness to accept vulnerability; reputation as socially accumulated evaluation inferred from visible cues), and HTMT ratios between these three constructs would be reported explicitly.

7.3 Embedded Factorial Experiment

Participants would evaluate identical self-presentation profiles within a standardized mock social-media platform. They would be randomly assigned to one of four authorship-labeling conditions: (a) a contextualized AI-assisted label specifying limited, human-supervised use; (b) a generic AI-generated label; (c) a human-authored label; and (d) no authorship information. Independently, they would be assigned to one of two platform-transparency environments: (a) high

transparency, featuring a clear policy summary, visible and consistently applied labels, and accessible explanations; or (b) low transparency, featuring minimal, inconsistently displayed, and poorly explained disclosure information. This 4×2 design aligns the platform-transparency construct with the same experimental environment in which the profile is evaluated, rather than asking participants to judge an unrelated platform. Content-Level Disclosure Quality would be operationalized through the contrast between the contextualized and generic AI labels, and the CDQ items would verify the perceived clarity, specificity, and explanatory detail of those labels. The PT items would verify the perceived transparency of the mock platform environment. Perceived AI Assistance would first serve as a manipulation check and would then be examined as a mediator of the label's effect on Perceived Authenticity (label → PAA → PAU), thereby connecting the manipulation to P1's perceptual mechanism. P8 would be tested through planned label contrasts, a PAA×CDQ interaction, and conditional indirect effects through PAA. P9 would be tested through the PAA×PT interaction and the high-versus-low transparency contrast. P10 would be evaluated through the PAA×AAL×CDQ interaction, with SPAIL and OAIK examined separately if the higher-order composite is not supported. To make reputation assessable from a single exposure, the profile stimuli would include visible social evaluation cues, including ratings, comments, follower counts, and recommendation indicators. These cues would be identical across conditions, moderate rather than extreme in valence, pretested for neutrality, and realistic for the selected platform, so that participants form a reputation judgment partly through authenticity rather than simply reproducing the displayed indicators. Causal claims would extend only to the manipulated label and platform-environment contrasts, and a debriefing step would disclose the constructed nature of the profiles, labels, and platform policies.

7.4 Data Analysis

Data would be analyzed in stages. Descriptive statistics and screening (missing data, outliers, skewness/kurtosis) would precede substantive analysis. Common method bias would be addressed first through procedural design, following established recommendations: predictor and outcome sections would be psychologically separated in the questionnaire, item order varied across respondents, anonymity guaranteed and stated, ambiguous and evaluative wording avoided, different response anchors used where justified, and the embedded experimental manipulation and cue-based reputation stimuli would provide measurement sources that do not depend solely on a single self-report style (Podsakoff et al., 2003, pp. 887–895). Statistical diagnostics, including Harman's single-factor test, together with marker-variable and collinearity-based tests, would then serve as supplementary verification rather than the primary remedy. Multicollinearity would be assessed through variance inflation factors (VIF < 5). For structural modeling, PLS-SEM is considered because the study's purpose is prediction-oriented theory development and because several measures are newly developed, which are conditions under which variance-based estimation is appropriate, not merely because the model contains moderation (Hair et al., 2019, pp. 4–7, 13–20). For a subsequent, strongly confirmatory study focused on

overall model fit with stabilized measures, covariance-based SEM may ultimately be more appropriate, and this two-stage logic (PLS-SEM for development, CB-SEM for confirmation) would be stated in advance. The structural model would be evaluated through path coefficients, R^2 , effect sizes f^2 , predictive relevance Q^2 , and bootstrapped significance testing with 5,000 resamples. P8 would be tested through planned contrasts between contextualized and generic AI labels, a PAA×CDQ interaction, and second-stage moderated-mediation analysis of the label→PAA→PAU pathway. P9 would be tested through the experimental high-versus-low platform-transparency contrast and a PAA×PT interaction with simple-slopes analysis. P10 would be tested through a three-way PAA×AAL×CDQ interaction, followed by conditional simple slopes; if the higher-order AAL composite is not supported, SPAIL and OAIK would be tested separately. Mediation (P7) would be tested through bootstrapped indirect effects, and experimental outcomes would also be compared using factorial ANCOVA across the eight 4×2 cells. Table 2 summarizes the plan.

Table 2. Proposed variable roles and analytic techniques.

Variable role	Constructs	Analytic technique
Independent	PAA, PAMV, PSS	EFA → CFA (split samples); structural paths to PAU (P1–P3)
Mediator	PAU	Bootstrapped indirect effects (P7)
Dependent	POR, IT	Structural paths (P4–P6); R^2 , Q^2
User-level moderator	AAL (or SPAIL and OAIK separately)	PAA×AAL×CDQ interaction and conditional simple slopes (P10)
Experimental boundary conditions	CDQ / authorship label (four levels); PT environment (high vs low)	Manipulation checks; planned contrasts; PAA×CDQ (P8); PAA×PT (P9); moderated mediation; 4×2 factorial ANCOVA
Controls	Age, platform type, usage intensity, prior AI-tool use	Theoretically justified paths only (see Section 7.4)

Control variables would not be entered wholesale as paths to the outcomes. Each control path would require theoretical justification, and several controls plausibly belong elsewhere in the model: prior AI-tool use may shape perceived AI assistance, platform type may shape perceived algorithmic mediation, usage intensity may shape awareness of platform transparency, and age may relate to AI literacy. The modeling default would therefore be to include a control path only where a specific confounding account exists, and to report models with and without controls, since unnecessary controls can distort rather than improve estimates. Ethical approval would be obtained from the relevant institutional review board prior to any data collection; informed consent, voluntary participation, anonymity, and the right to withdraw would be guaranteed and stated in the instrument.

8. Theoretical Contributions

The AMA model makes three contributions to communication and information science. First, it supplies a construct-level bridge between research traditions that have remained separate despite conceptual adjacency: it embeds the authorship question raised by AI-mediated communication within the audience-evaluation apparatus of credibility research (Hancock et al., 2020, pp. 89–92; Ohanian, 1990, pp. 41–46; Sundar, 2008, pp. 73–81, 84–86), while incorporating the visibility beliefs theorized by algorithmic power scholarship and documented by algorithm-awareness research as a distinct antecedent rather than a background condition (Bucher, 2012, p. 1164; Bucher, 2017, pp. 30–35, 40–42; Gran et al., 2021, pp. 1781–1784, 1790–1792). The model's central analytic distinction, between the perceived authorship of digital identity content and the algorithmic mechanisms governing its visibility, was not, in our review, formalized in any existing framework.

Second, the model extends the Replicant Effect from an isolated experimental finding into a nomological network (Jakesch et al., 2019, pp. 1–2, 8–10): it specifies three analytically distinct boundary conditions, content-level disclosure quality, platform transparency, and audience AI literacy, and the downstream pathway (authenticity → reputation → trust) through which the effect should propagate. Third, methodologically, the model and its accompanying preliminary item pool (Appendix) offer a replicable template through which the propositions can be converted into hypotheses and tested across platforms, cultures, and regulatory environments, including non-Western contexts where the AI-MC evidence base is currently thin.

9. Practical and Societal Implications

For platform designers and governance teams, the model separates two decisions that are often conflated. Content-level disclosure quality concerns what a specific label tells the audience about the degree of AI involvement and human oversight, while platform transparency concerns whether such labels are visible, understandable, and consistently applied across the service. If P8 holds, contextualized labels should preserve authenticity better than generic AI-generated warnings. If P9 holds, a coherent platform-level disclosure environment should improve audience calibration beyond the effect of any single label. Poorly contextualized or inconsistently applied labels can instead trigger assumptions of complete machine authorship and reduce trust (Altay & Gilardi, 2024, pp. 1–3, 7–8; Lin & Zhang, 2026, pp. 1–3, 14–17; Schilke & Reimann, 2025, pp. 1–2, 12–14; Shi & Jiang, 2026, pp. 1, 5, 14–16). For regulators implementing obligations such as Article 50 of the European Union Artificial Intelligence Act, the model identifies two empirical questions: whether labels accurately communicate the degree of AI involvement and whether platforms apply them consistently (European Union, 2024, Art. 50). For AI developers, perceived AI assistance functions as a design target because defaults about when and how tools disclose their involvement shape audience inference. For educators and digital-literacy policymakers, P10 suggests that literacy should distinguish self-perceived familiarity from objectively demonstrated knowledge and should train users to differentiate assisted, edited, and fully generated content. Greater knowledge may increase tolerance of

proportionate use but also sharpen sanctions against concealed use (Jakesch et al., 2023, pp. 1–2, 6–8; Long & Magerko, 2020, pp. 2–7; Ng et al., 2021, pp. 2–8). For professionals, entrepreneurs, and public communicators, the model articulates a practical trade-off: AI assistance offers production efficiency, but its authenticity cost depends on the degree of involvement, the quality of disclosure, the platform environment, and audience expectations.

10. Limitations and Future Research

As a conceptual article, the central limitation is inherent: the propositions await empirical validation through the strategy specified in Section 7, and the model's value will ultimately be decided by data. The literature base, assembled through an integrative rather than systematic review (Section 2), bounds the article's gap claims to the search strategy described there. Several design limitations of the proposed validation strategy should also be acknowledged in advance. Self-report measurement of most constructs invites common-method concerns despite the mitigation procedures specified; future work should incorporate behavioral or trace-data indicators of actual, not merely perceived, AI assistance where platform access permits. The proposed cross-sectional design cannot capture how authenticity penalties evolve as generative tools normalize; longitudinal designs tracking the co-evolution of AI literacy and authenticity judgments would address this, particularly since training effects on detection appear limited (Jakesch et al., 2023, pp. 1–2, 6–8).

The model itself is deliberately parsimonious, and several extensions were explicitly deferred. The directional form of P8–P10 is intentionally conditional: recent evidence indicates that disclosure and literacy can generate different outcomes depending on label specificity, human oversight, content stakes, prior attitudes, and whether AI involvement is concealed or voluntarily disclosed (Lin & Zhang, 2026, pp. 1–3, 14–17; Schilke & Reimann, 2025, pp. 1–2, 12–14; Shi & Jiang, 2026, pp. 1, 5, 14–16). Future studies should therefore estimate content-level, platform-level, and user-level conditional effects separately rather than assuming that transparency or literacy is uniformly protective. The proposed higher-order specification of AAL also requires empirical justification; self-perceived literacy and objective knowledge may function as distinct moderators rather than forming one composite. Perceived strategic self-presentation is specified as unipolar; a bipolar authentic-versus-strategic construct would require separate items for authentic-disclosive presentation. Source credibility is currently the theoretical anchor of perceived authenticity rather than a separate construct; future elaborations could model expertise, trustworthiness, and attractiveness as a distinct mediating layer, with corresponding measurement items and structural paths (Ohanian, 1990, pp. 41–46). Similarly, a direct path from perceived algorithmic mediation to trust was excluded pending evidence on the P2 relationship; if algorithmic-mediation beliefs prove to affect trust independently of authenticity, the model should be extended accordingly. Finally, the framework should be tested beyond individual identities, including organizational and institutional digital identities, and across regulatory environments with markedly different transparency regimes.

11. Conclusion

Generative artificial intelligence has not created the problem of curated online identity; self-presentation research has addressed that problem for decades (Goffman, 1959, pp. 17–25). What it has introduced is a genuinely new variable: audience uncertainty about who, or what, authored the performance. This article has argued that the scholarly resources for understanding this condition exist but remain fragmented across self-presentation theory, algorithmic visibility research, human–machine communication, and credibility scholarship. Existing experimental evidence already establishes both that perceived AI assistance measurably affects trust (Jakesch et al., 2019, pp. 1–2, 8–10) and that audiences cannot reliably detect it unaided (Jakesch et al., 2023, pp. 1–2, 6–8; Köbis & Mossink, 2021, pp. 1–2, 7–9). The Algorithmically Mediated Authenticity model integrates these traditions into a single conceptual structure organized around one distinction: between the perceived authorship of identity content and the algorithmic mechanisms governing its visibility. Ten research propositions specify the relationships the model asserts, a complete survey-experimental strategy specifies how those propositions can be converted into hypotheses and tested, and a preliminary item pool is proposed for future validation. The framework is offered as a foundation: its propositions are falsifiable, its measurement starting point is explicit about what is adapted and what is new, and its validation in Western and non-Western contexts alike is the necessary next step for a communication science confronting an environment in which human and algorithmic authorship can no longer be assumed apart.

References

- 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>
- boyd, d. (2014). *It's complicated: The social lives of networked teens*. Yale University Press.
- Bucher, T. (2012). Want to be on top? Algorithmic power and the threat of invisibility on Facebook. *New Media & Society*, 14(7), 1164–1180. <https://doi.org/10.1177/1461444812440159>
- Bucher, T. (2017). The algorithmic imaginary: Exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society*, 20(1), 30–44. <https://doi.org/10.1080/1369118X.2016.1154086>
- Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics* (pp. 7282–7296). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.acl-long.565>
- Coalition for Content Provenance and Authenticity. (n.d.). C2PA technical specification. Retrieved July 2026, from <https://c2pa.org/specifications/>
- Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science*, 49(10), 1407–1424. <https://doi.org/10.1287/mnsc.49.10.1407.17308>
- European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial

intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>

Fouquaert, T., & Mechant, P. (2022). Making curation algorithms apparent: A case study of 'Instawareness' as a means to heighten awareness and understanding of Instagram's algorithm. *Information, Communication & Society*, 25(12), 1769–1789. <https://doi.org/10.1080/1369118X.2021.1883707>

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18(1), 39–50. <https://doi.org/10.1177/002224378101800104>

Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), *Media technologies: Essays on communication, materiality, and society* (pp. 167–194). MIT Press.

Goffman, E. (1959). *The presentation of self in everyday life*. Doubleday.

Gran, A.-B., Booth, P., & Bucher, T. (2021). To be or not to be algorithm aware: A question of a new digital divide? *Information, Communication & Society*, 24(12), 1779–1796. <https://doi.org/10.1080/1369118X.2020.1736124>

Guzman, A. L. (2019). Voices in and of the machine: Source orientation toward mobile virtual assistants. *Computers in Human Behavior*, 90, 343–350. <https://doi.org/10.1016/j.chb.2018.08.009>

Guzman, A. L., & Lewis, S. C. (2020). Artificial intelligence and communication: A Human-Machine Communication research agenda. *New Media & Society*, 22(1), 70–86. <https://doi.org/10.1177/1461444819858691>

Hair, J. F., Risher, J. J., Sarstedt, M., & Ringle, C. M. (2019). When to use and how to report the results of PLS-SEM. *European Business Review*, 31(1), 2–24. <https://doi.org/10.1108/EBR-11-2018-0203>

Hancock, J. T., Naaman, M., & Levy, K. (2020). AI-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication*, 25(1), 89–100. <https://doi.org/10.1093/jcmc/zmz022>

Hargittai, E., Gruber, J., Djukaric, T., Fuchs, J., & Brombach, L. (2020). Black box measures? How to study people's algorithm skills. *Information, Communication & Society*, 23(5), 764–775. <https://doi.org/10.1080/1369118X.2020.1713846>

Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, 43(1), 115–135. <https://doi.org/10.1007/s11747-014-0403-8>

Jakesch, M., French, M., Ma, X., Hancock, J. T., & Naaman, M. (2019). AI-mediated communication: How the perception that profile text was written by AI affects trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). ACM. <https://doi.org/10.1145/3290605.3300469>

Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11), e2208839120. <https://doi.org/10.1073/pnas.2208839120>

Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior*, 114, 106553. <https://doi.org/10.1016/j.chb.2020.106553>

Lehman, D. W., O'Connor, K., Kovács, B., & Newman, G. E. (2019). Authenticity. *Academy of Management Annals*, 13(1), 1–42. <https://doi.org/10.5465/annals.2017.0047>

- Lin, T., & Zhang, Y. (2026). Visible sources and invisible risks: Exploring the impact of AI disclosure on perceived credibility of AI-generated content. *Journal of Science Communication*, 25(1), A09. <https://doi.org/10.22323/358020260107085703>
- Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). Association for Computing Machinery. <https://doi.org/10.1145/3313831.3376727>
- Marwick, A. E., & boyd, d. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, 13(1), 114–133. <https://doi.org/10.1177/1461444810365313>
- Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2, 100041. <https://doi.org/10.1016/j.caeai.2021.100041>
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.
- Ohanian, R. (1990). Construction and validation of a scale to measure celebrity endorsers' perceived expertise, trustworthiness, and attractiveness. *Journal of Advertising*, 19(3), 39–52. <https://doi.org/10.1080/00913367.1990.10673191>
- Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5), 879–903. <https://doi.org/10.1037/0021-9010.88.5.879>
- Schilke, O., & Reimann, M. (2025). The transparency dilemma: How AI disclosure erodes trust. *Organizational Behavior and Human Decision Processes*, 188, 104405. <https://doi.org/10.1016/j.obhdp.2025.104405>
- Shi, Y., & Jiang, Z. (2026). Consumer responses to AI disclosure labels: The role of novelty and authenticity. *SAGE Open*, 16(1), 1–18. <https://doi.org/10.1177/21582440261417793>
- Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. In M. J. Metzger & A. J. Flanagin (Eds.), *Digital media, youth, and credibility* (pp. 73–100). MIT Press.
- Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, 6(1), 1–13. <https://doi.org/10.1177/2056305120903408>

Appendix: Preliminary Item Pool for Future Validation

All items would be measured on 7-point Likert scales (1 = strongly disagree, 7 = strongly agree), except the objective AI-knowledge items, which would be scored for accuracy. The pool is preliminary and not yet validated. Table A1 lists items adapted from, or theoretically informed by, established measures. Table A2 lists newly developed items, including the two lower-order dimensions proposed for Audience AI Literacy. The CDQ and PT items are manipulation-check items: they verify whether participants perceived the contextualized label and high-transparency platform environment as clearer, more specific, more visible, and more consistent than their comparison conditions. All new items would undergo the full content, construct, convergent, discriminant, and criterion validation sequence described in Section 7.2 before substantive use.

Table A1. Items adapted from or theoretically informed by established measures.

Construct	Proposed items (illustrative)	Adapted from / theoretically informed by
Perceived AI Assistance (PAA)	PAA1. I believe AI tools were used to create this content. PAA2. This content shows signs of AI generation or substantial AI modification. PAA3. Parts of this profile appear to have been generated or substantially modified by AI. PAA4. AI appears to have played a meaningful role in producing this identity presentation.	AI-perception manipulation-check items (Jakesch et al., 2019, pp. 1–2, 8–10); AI-MC framework (Hancock et al., 2020, pp. 89–92)
Perceived Authenticity (PAU)	PAU1. This content genuinely reflects who this person is. PAU2. This feels like an authentic expression of this person's voice. PAU3. This person comes across as real rather than manufactured.	Trustworthiness dimension of source credibility (Ohanian, 1990, pp. 41–46)
Perceived Online Reputation (POR)	POR1. Based on the ratings and comments shown, this person is well-regarded by others. POR2. This person appears to have a good reputation among their audience. POR3. Others seem to value this person's content.	Online feedback and reputation mechanisms (Dellarocas, 2003, pp. 1407–1411); expertise dimension of source credibility (Ohanian, 1990, pp. 41–46); social evaluation cues embedded in stimuli (Section 7.3)
Interpersonal Trust (IT)	IT1. I would rely on this person for what they offer (e.g., a booking, a purchase, a collaboration). IT2. I am confident in the accuracy of the information this person provides. IT3. I would be willing to accept some risk (e.g., share information, make a payment) based on this profile. IT4. This person seems reliable and predictable.	Perceived-trustworthiness measure (Jakesch et al., 2019, pp. 1–2, 8–10), reworded toward reliance, confidence, vulnerability, and reliability

Table A2. Newly developed items (to be validated).

Construct	Proposed items (illustrative)	Developed from
Perceived Algorithmic Mediation of Visibility (PAMV)	PAMV1. Algorithms determine who sees this person's posts. PAMV2. This person's visibility depends on platform ranking systems. PAMV3. This person's prominence reflects algorithmic promotion more than audience choice.	Algorithmic power and visibility reasoning (Bucher, 2012, pp. 1164–1167, 1173–1175; Bucher, 2017, pp. 30–35, 40–42); algorithm awareness research (Gran et al., 2021, pp. 1781–1784, 1790–1792; Hargittai et al., 2020, pp. 765–769)
Perceived Strategic Self-Presentation (PSS)	PSS1. This person presents an idealized version of themselves. PSS2. This person's posts feel like a personal brand. PSS3. This person's self-presentation seems calculated for effect.	Micro-celebrity and self-branding accounts (Marwick & boyd, 2011, pp. 114–119)
Content-Level Disclosure Quality (CDQ), manipulation-check items	CDQ1. The label clearly explains whether AI assisted or fully generated the content. CDQ2. The label makes the role of human oversight clear. CDQ3. The label provides enough context to interpret the degree of AI involvement. These items verify that the contextualized label is perceived as clearer and more informative than the generic label.	AI-disclosure and labeling research (Altay & Gilardi, 2024, pp. 1–3, 7–8; Lin & Zhang, 2026, pp. 1–3, 14–17; Schilke & Reimann, 2025, pp. 1–2, 12–14; Shi & Jiang, 2026, pp. 1, 5, 14–16)
Platform Transparency (PT), manipulation-check items	PT1. The experimental platform clearly labels content created or assisted by AI. PT2. AI-related labels in the experimental platform are easy to notice and understand. PT3. The experimental platform applies its AI-labeling rules consistently. These items verify the high-versus-low transparency manipulation.	Content-provenance and disclosure frameworks (C2PA, n.d., "Technical Specifications" section); EU AI Act transparency logic (European Union, 2024, Art. 50)
Self-Perceived AI Literacy (SPAIL), lower-order AAL dimension	SPAIL1. I understand how generative AI tools work. SPAIL2. I know what kinds of content current AI tools can produce. SPAIL3. I keep up with developments in AI content generation.	AI literacy competency frameworks (Long & Magerko, 2020, pp. 2–7; Ng et al., 2021, pp. 2–8)
Objective AI Knowledge (OAIK), lower-order AAL dimension	OAIK1. Identify which listed capabilities current generative tools realistically possess. OAIK2. Distinguish content generation from content recommendation in described scenarios. OAIK3. Correctly interpret the meaning of a platform AI label. OAIK4. Recognize that fluent, confident AI output does not guarantee factual accuracy.	AI literacy competency frameworks (Long & Magerko, 2020, pp. 2–7; Ng et al., 2021, pp. 2–8)

