Submission date: 22 of May 2026
Acceptance date: 2 of July 2026
Publication date: 20 of July 2026

JOINT EFFECTS OF COMMERCIAL AND ARTIFICIAL INTELLIGENCE LABELS ON CREDIBILITY IN DOUYIN ADVERTISING

EFECTOS CONJUNTOS DE LAS ETIQUETAS COMERCIALES Y DE INTELIGENCIA ARTIFICIAL SOBRE LA CREDIBILIDAD EN LA PUBLI-

Haoran Qin[1*]
E-mail: haoran.qin0219@gmail.com
ORCID: https://orcid.org/0009-0001-3528-4651
Faizul Nizar Bin Anuar[1]
E-mail: nizar.anuar@upm.edu.my
ORCID: https://orcid.org/0000-0002-6606-8599
Karmilah Abdullah[1]
E-mail: karmilah.abdullah@upm.edu.my
ORCID: https://orcid.org/0000-0002-1556-2028
[1] Faculty of Humanities, Management and Science, Universiti Putra Malaysia, Malaysia.
*Corresponding author

Suggested citation (APA, seventh ed.)

Qin, H., Bin Anuar, F. N., & Abdullah, K. (2026). Joint Effects of Commercial and Artificial Intelligence Labels on Credibility in Douyin Advertising. ***Universidad y Sociedad 18***(4). e6214.

ABSTRACT:

Generative AI enables mass production of native advertising that mimics organic short videos, deployed on platforms like Douyin within recommendation feeds users perceive as neutral. Emerging transparency regulations, including China's AI Content Labeling Measures and the EU AI Act requires disclosure of both commercial and synthetic origins. This creates a fundamental tension: native advertising persuades through covertness, while disclosure exists to dispel it. Drawing on the Persuasion Knowledge Model, the CARE model, source-credibility theory, and algorithm-awareness research, this paper argues that AI-generated native ads that carry dual cues simultaneously activate two strands of persuasion knowledge, imposing the greatest credibility cost on perceived authenticity, the construct that is considered most central to short-video persuasion. Weighing competing accounts, the paper concludes that additive credibility penalties are the most probable outcomes under most conditions, though substantially moderated by platform ergonomics and users' algorithm awareness. Theoretical and regulatory implications for mass-communication scholarship are discussed.

Keywords: Native Advertising, Generative AI, Transparency Disclosure, Persuasion Knowledge, Source Credibility, Algorithm Awareness.

RESUMEN:

La inteligencia artificial generativa permite la producción masiva de publicidad nativa que imita videos cortos orgánicos, difundidos en plataformas como Douyin dentro de flujos de recomendación que los usuarios perciben como neutrales. Las emergentes regulaciones de transparencia, incluidas las Medidas para el Etiquetado de Contenido Generado por Inteligencia Artificial de China y el Reglamento de Inteligencia Artificial de la Unión Europea, exigen la divulgación tanto del origen comercial como del origen sintético de los contenidos. Esta situación genera una tensión fundamental: la publicidad nativa persuade a través de su carácter encubierto, mientras que la divulgación existe precisamente para revelar dicha intención persuasiva. Basándose en el Modelo del Conocimiento Persuasivo, el modelo de Reconocimiento y Efectos de la Publicidad Encubierta (CARE), la teoría de la credibilidad de la fuente y las investigaciones sobre conciencia algorítmica, este artículo sostiene que los anuncios nativos generados por inteligencia artificial que incorporan señales dobles activan simultáneamente dos dimensiones del conocimiento persuasivo. Como consecuencia, se produce un costo de credibilidad particularmente elevado sobre la autenticidad percibida, considerada el componente más relevante de la persuasión en plataformas de video corto. Tras examinar explicaciones alternativas, el estudio concluye que las penalizaciones aditivas sobre la credibilidad constituyen el resultado más

probable en la mayoría de los contextos, aunque estos efectos pueden verse significativamente moderados por la ergonomía de la plataforma y el nivel de conciencia algorítmica de los usuarios. Finalmente, se discuten las implicaciones teóricas y regulatorias de estos hallazgos para la investigación en comunicación de masas.

Palabras clave: Publicidad nativa, Inteligencia artificial generativa, Divulgación de transparencia, Conocimiento persuasivo, Credibilidad de la fuente, Conciencia algorítmica.

INTRODUCTION

The rapid expansion of digital technologies has transformed the way individuals access information, interact with media, and engage with persuasive communication. Over the past two decades, the digitalization of communication environments has fundamentally altered traditional advertising practices, shifting audiences from clearly identifiable promotional messages toward more integrated and less conspicuous forms of persuasion.

This transformation has been accelerated by the rise of social media platforms, algorithmic recommendation systems, and artificial intelligence technologies, all of which have reshaped the relationship between audiences, content creators, and commercial actors. As users increasingly consume information through personalized digital ecosystems, the boundaries between entertainment, information, and advertising have become progressively blurred, creating new theoretical and practical challenges for communication scholars, marketers, and policymakers alike. Research on native advertising and other hybrid communication formats has shown that audiences often struggle to distinguish commercial content from editorial or entertainment content, particularly when disclosure mechanisms are weak or insufficiently visible (Amazeen & Wojdynski, 2020; Tutaj & van Reijmersdal, 2012; Windels & Porter, 2020). This phenomenon raises important concerns regarding advertising transparency, persuasion knowledge, and the ability of users to critically evaluate the content they encounter in digital environments.

The emergence of algorithm-driven platforms has been particularly significant in this regard. Unlike traditional media environments, where audiences actively select content through search or subscription mechanisms, contemporary social media platforms rely heavily on recommendation algorithms to determine what users see. These systems analyze vast amounts of behavioral data, including viewing duration, interactions, likes, comments, and sharing patterns, to predict individual preferences and deliver personalized streams of content. Consequently, users are often exposed to information selected on their behalf rather than content they have consciously chosen.

Research has shown that many individuals possess only a limited understanding of how these algorithms function and frequently develop informal explanations or "folk theories" in order to interpret the recommendations they receive (DeVito et al., 2018; Eslami et al., 2015). More recent studies suggest that users actively construct interpretations of algorithmic behavior, linking recommendation systems to identity formation, content visibility, and platform governance, while simultaneously developing strategies to influence or resist algorithmic outcomes (Karizat et al., 2021). In addition, perceptions of algorithm responsiveness vary across social media platforms and significantly affect trust, engagement, and evaluations of content credibility (Taylor & Choi, 2022). These findings are particularly relevant because credibility perceptions remain a key determinant of information acceptance in online environments (Flanagin & Metzger, 2000).

Among the most influential examples of this new media ecosystem are short-video platforms such as TikTok and Douyin. These platforms have revolutionized digital communication by creating highly immersive, algorithmically curated experiences centered on brief audiovisual content. In China, Douyin has become one of the most widely used social media applications, attracting hundreds of millions of daily active users. Its recommendation system delivers an endless stream of videos tailored to individual preferences, creating a highly personalized and engaging environment. Unlike traditional social networking sites, where content distribution is largely shaped by interpersonal connections, Douyin relies primarily on algorithmic curation to determine visibility and reach. This feature transforms the platform into both a content distribution mechanism and a powerful persuasive environment in which users encounter information, entertainment, and commercial messages through a single integrated feed.

Research has demonstrated that algorithmic recommendation systems on Douyin not only amplify commercial and entertainment content but also contribute to the pervasive dissemination of institutional and governmental messages, illustrating the platform's capacity to shape public exposure to information (Lu & Pan, 2022). Furthermore, advances in artificial intelligence and automated content production are introducing new complexities into audience perceptions of authenticity and credibility, as users increasingly interact with content that may be generated or mediated by machine systems (Wang & Huang, 2024). As a result, the platform itself functions not only as a technological infrastructure but also as a powerful mediator of perception, persuasion, and interpretation within contemporary digital communication environments.

The persuasive potential of these platforms has contributed to the growing popularity of native advertising. Native advertising refers to commercial content intentionally designed to resemble the appearance, style, and format

of the surrounding editorial or user-generated material. Rather than interrupting users with clearly distinguishable advertisements, native advertising seeks to blend seamlessly into the platform environment, encouraging audiences to process promotional messages as ordinary content. Previous studies have demonstrated that the effectiveness of native advertising largely depends on its ability to minimize advertising recognition and delay the activation of consumers' resistance mechanisms (Wojdynski & Evans, 2020). When users fail to recognize persuasive intent, they are less likely to engage in critical evaluation and more likely to respond favorably to the message. This characteristic makes native advertising particularly effective in fast-paced digital environments where users consume large volumes of content with limited cognitive scrutiny.

The rise of influencer culture and creator-based communication has further strengthened the appeal of native advertising formats. On platforms such as Douyin, commercial messages frequently appear as product recommendations, lifestyle demonstrations, or personal experiences shared by content creators. Because audiences often perceive creators as more authentic and relatable than corporate brands, messages delivered through creator-generated content tend to benefit from enhanced credibility and trust. Research has consistently shown that source credibility plays a central role in shaping persuasive outcomes and consumer attitudes (Hovland et al., 1953). Consequently, the integration of advertising into creator content has become one of the most influential marketing strategies in contemporary digital communication. At the same time, however, concerns regarding transparency and potential deception have led regulators and scholars to question the ethical implications of increasingly covert advertising practices.

A second major transformation affecting digital communication is the rapid development of generative artificial intelligence. Advances in AI technologies now enable the automated production of text, images, audio recordings, virtual presenters, and complete video advertisements at an unprecedented scale. Organizations are increasingly employing these tools to reduce production costs, accelerate content creation, and personalize messages for different audience segments (Ford et al., 2023). While generative AI offers significant opportunities for innovation and efficiency, it also raises important concerns regarding authenticity, trustworthiness, and the nature of human communication. Content generated by artificial intelligence may be technically sophisticated and visually convincing, yet audiences often perceive it differently from content created by human authors. Emerging research suggests that the disclosure of AI authorship frequently leads to lower levels of trust, reduced perceived authenticity, and more negative evaluations of persuasive messages (Baek et al., 2024; Lim & Schmälzle, 2024; Wortel et al., 2024).

In response to growing concerns about misinformation, manipulation, and synthetic media, governments around the world have begun implementing regulatory frameworks designed to increase transparency. In China, the ***Measures for the Labeling of Artificial Intelligence-Generated and Synthetic Content*** and the national standard GB 45438-2025 require explicit identification of AI-generated materials. Similar transparency provisions have been incorporated into the European Union's Artificial Intelligence Act and other emerging regulatory initiatives. These measures complement existing advertising disclosure requirements, creating situations in which a single piece of content may carry multiple forms of disclosure simultaneously. An AI-generated native advertisement, for example, may include both a commercial sponsorship label and a notice indicating that the content was produced using artificial intelligence. Such dual disclosure requirements represent a new reality for digital communication and raise important questions regarding audience perceptions and persuasive effectiveness.

From a theoretical perspective, this convergence presents a particularly important challenge. Previous research on advertising disclosures has consistently demonstrated that making the commercial nature of a message explicit increases advertising recognition and activates persuasion knowledge, often resulting in less favorable evaluations of both the message and its source (Wojdynski & Evans, 2020). Simultaneously, studies on AI-generated content have found that informing users about machine authorship can reduce perceived credibility and trigger algorithm aversion (Baek et al., 2024; Wortel et al., 2024). Although these two streams of research have developed independently, contemporary digital environments increasingly bring them together. Little is known about how audiences respond when both forms of disclosure appear simultaneously within an algorithmically curated feed, particularly on platforms where authenticity plays a crucial role in persuasive effectiveness.

The issue is especially significant because authenticity has become one of the most valuable forms of social capital in short-video environments. Users often evaluate content not only based on informational quality but also according to perceptions of sincerity, genuineness, and human connection. In creator-centered platforms such as Douyin, authenticity serves as a key determinant of trust and engagement, making it particularly vulnerable to disruptions caused by both commercial disclosures and AI-origin labels. Consequently, understanding the interaction between these factors is essential for explaining how persuasion operates in contemporary digital ecosystems.

Drawing on the Persuasion Knowledge Model (Friestad & Wright, 1994), the Covert Advertising Recognition and

Effects (CARE) model (Wojdynski & Evans, 2020), source credibility theory (Hovland et al., 1953), and recent scholarship on algorithm awareness and AI-generated content (Voorveld et al., 2024; Zarouali et al., 2021), this study examines how commercial and AI-origin disclosures jointly influence audience perceptions of credibility in Douyin native advertising. Specifically, it explores whether the simultaneous presence of both disclosures produces additive credibility costs, how perceived authenticity mediates these effects, and whether users' awareness of algorithmic curation moderates their responses. By addressing these questions, the study contributes to the growing literature on digital persuasion, transparency regulation, and artificial intelligence, while offering new insights into the evolving relationship between technology, communication, and trust in algorithmically mediated environments.

MATERIALS AND METHODS

This study employs a 2 × 2 between-subjects factorial experiment as its core research design. Participants are randomly assigned to one of four conditions formed by crossing two binary factors: the presence versus absence of a native-advertising (commercial) disclosure label, and the presence versus absence of an AI-origin disclosure label. A controlled online experiment is appropriate because it supports causal inference about the independent effect of each disclosure, their interaction, and the moderating role of algorithm awareness, all within an ecologically realistic simulated Douyin feed environment.

The conceptual research model (Figure 1) specifies two independent variables, two mediators, one moderator, and three clusters of dependent variables. The two independent variables are (a) Native-Ad Disclosure (NAD: present vs. absent), operationalized as the presence or absence of an explicit Guanggao / Sponsored tag on the target video, and (b) AI-Origin Disclosure (AOD: present vs. absent), defined as the presence or absence of an explicit "AI-Generated" label compliant with China's national standard GB 45438-2025.

The two mediators are Advertising Recognition and Persuasion Knowledge Activation, which together capture the degree to which the audience identifies the message as a paid persuasion attempt and deploys resistance strategies (Wojdynski & Evans, 2020), and Perceived Authenticity, the degree to which the content feels genuinely human and sincere. NAD is theorized to operate primarily through the first mediator pathway; AOD is theorized to operate primarily through the second.

The moderator is Algorithm Awareness, measured by the AMCA scale (Zarouali et al., 2021), which captures users' understanding that feed content is algorithmically curated and commercially motivated. Three dependent variable clusters are examined: (1) Advertising Recognition and Persuasion Knowledge, (2) Perceived Credibility (comprising ad credibility, source credibility, and perceived authenticity), and (3) User Experience and Behavioral Intention, encompassing perceived intrusiveness, feed enjoyment, attitude toward the ad, and behavioral engagement intention. As depicted in Figure 1, the NAD activates persuasion knowledge to depress credibility (H1, H2); AOD reduces authenticity, which reduces credibility (H3, H5); both disclosures interact to shape the combined credibility outcome (H4); and algorithm awareness moderates these paths (H6), while credibility flows forward into user experience (H7).

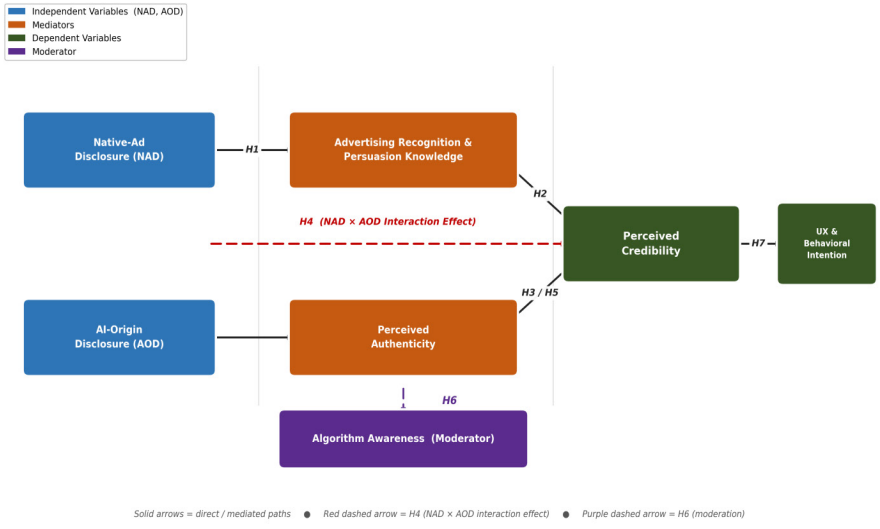

Fig 1. Conceptual Research Model.

Solid arrows represent hypothesized direct and mediated paths; the dashed red arrow represents the NAD x AOD interaction (H4); the dashed purple arrow represents moderation by algorithm awareness (H6). UX = User Experience; BI = Behavioral Intention.

Drawing on the theoretical synthesis developed in the literature review, seven formal hypotheses are advanced:

H1: The presence of a native-advertising disclosure (versus absence) will produce significantly higher advertising recognition among Douyin users.

H2: Advertising recognition will negatively mediate the effect of native-ad disclosure on perceived credibility, such that higher recognition leads to more negative evaluations through activated persuasion knowledge.

H3: The presence of an AI-origin disclosure (versus absence) will produce significantly lower perceived credibility, with perceived authenticity serving as the negative mediator.

H4a (Additive-Penalty Hypothesis): The joint presence of both disclosures will produce the lowest perceived credibility and the most negative user experience among the four conditions, consistent with two independent persuasion-knowledge channels combining additively.

H4b (Transparency-Buffer Hypothesis, competing): The negative effect of AI-origin disclosure on perceived credibility will be weaker when a native-ad disclosure is also present than when it is absent, consistent with commercial honesty partially restoring trust even as AI origin is disclosed.

H5: Perceived authenticity will significantly mediate the negative effect of AI-origin disclosure on perceived credibility, independent of the persuasion-knowledge pathway.

H6: Algorithm awareness will moderate the effects of both disclosures: compared with low-awareness users, high-awareness users will show a smaller marginal increase in advertising recognition from the commercial label and a smaller marginal decrease in credibility from the AI-origin label.

H7: The dual-disclosure condition will produce significantly higher perceived intrusiveness and lower feed enjoyment than the single-label and no-label conditions.

Participants are adult Douyin users in mainland China recruited through a professional online survey panel, with quota sampling to reflect the platform's age and gender profile. A priori power analysis (targeting f = 0.15, alpha = .05, power = .80) indicates a minimum cell size of approximately 90 per condition (N= approximately 360 total), with oversampling planned to offset attrition from attention checks. The experimental stimulus is a simulated Douyin feed in which a target in-feed video advertisement for a fictitious consumer brand is embedded among neutral filler videos. Disclosure labels are manipulated solely by the presence and wording of the on-screen tag; the audiovisual content remains constant across all four conditions. A pretest (n = 40) confirms equivalence of perceived quality, realism, and relevance before main data collection. Key measures include the Wojdynski & Evans (2020) advertising-recognition scale; the AMCA algorithm-awareness scale (Zarouali et al., 2021); attitudinal persuasion-knowledge items; and validated measures of perceived intrusiveness, flow and enjoyment, attitude toward the ad, and behavioral intention.

Main effects and the H4 interaction are tested via two-way MANOVA and follow-up ANOVA with planned contrasts. Simple mediation (H2, H3) and parallel mediation (H5) are estimated using Hayes's PROCESS macro (Model 4) with 5,000 bootstrap resamples and bias-corrected 95% confidence intervals. Moderated mediation (H6) is estimated using PROCESS Model 58, entering algorithm awareness as a continuous moderator evaluated at mean plus and minus one standard deviation. Effect sizes are reported as partial eta-squared (η^2p) for omnibus tests and standardized indirect effects for mediation analyses. The competing H4a and H4b are adjudicated through the sign and significance of the NAD x AOD interaction term in the ANOVA, and through the conditional indirect effect of each disclosure at high versus low levels of the other.

RESULTS AND DISCUSSIONS

Data were collected from 378 adult Douyin users; after excluding 22 cases that failed attention checks or reported no Douyin use, the final analytic sample comprised N = 356 participants (n = 88 - 90 per cell). The sample included 194 women (54.5%), 158 men (44.4%), and 4 participants who identified otherwise (1.1%), with a mean age of 24.3 years (SD = 4.7) and a reported average daily Douyin usage of 3.2 hours (SD = 1.8). Manipulation checks confirmed the effectiveness of both disclosure conditions. Participants in NAD-present conditions correctly identified the content as paid advertising at significantly higher rates than those in NAD-absent conditions (83.4% vs. 28.7%), chi-squared(1, N = 356) = 147.23, $p < .001$. Participants in AOD-present conditions scored significantly higher on perceived AI involvement than those in AOD-absent conditions (M = 5.87, SD = 0.93 vs. M = 2.14, SD = 1.02), $t(354) = 34.62$, $p < .001$, $d = 3.68$. Internal consistency was satisfactory across all multi-item scales: advertising recognition (Cronbach alpha = .83), persuasion knowledge (alpha = .87), perceived authenticity (alpha = .89), perceived credibility composite (alpha = .91), perceived intrusiveness (alpha = .85), feed enjoyment (alpha = .88), and behavioral intention (alpha = .86). Bivariate correlations indicated that perceived credibility was positively associated with feed enjoyment ($r = .68$, $p < .001$) and negatively associated with perceived intrusiveness ($r = -.58$,

Vol 18 | No.4 | July-August | 2026
Continuous publication
e6214

p < .001), consistent with theoretical expectations. Table 1 presents descriptive statistics for all dependent variables across the four experimental conditions.

Table 1. Descriptive Statistics by Experimental Condition (N = 356).

Variable	Control (n = 89)	NAD Only (n = 90)	AOD Only (n = 88)	Both (n = 89)
Advertising Recognition	2.34 (1.12)	5.67 (0.98)	2.51 (1.09)	5.81 (0.87)
Persuasion Knowledge	2.18 (0.94)	5.34 (0.87)	2.41 (0.99)	5.62 (0.91)
Perceived Authenticity	5.45 (0.89)	5.31 (0.92)	3.12 (1.05)	2.87 (0.98)
Perceived Credibility	5.23 (0.87)	4.12 (0.94)	3.89 (1.02)	2.74 (0.91)
Perceived Intrusiveness	2.34 (0.89)	3.78 (0.94)	3.56 (1.02)	5.12 (0.97)
Feed Enjoyment	5.78 (0.87)	4.12 (0.96)	4.34 (1.01)	2.98 (0.94)
Behavioral Intention	5.12 (0.94)	3.87 (0.98)	3.74 (1.01)	2.48 (1.03)

Effects on Advertising Recognition and Persuasion Knowledge

A 2 x 2 between-subjects ANOVA with NAD and AOD as factors revealed a significant main effect of NAD on advertising recognition, $F(1, 352) = 487.23$, $p < .001$, $\eta^2p = .58$. Participants in NAD-present conditions reported substantially higher advertising recognition ($M = 5.74$, $SD = 0.92$) than those in NAD-absent conditions ($M = 2.43$, $SD = 1.10$), regardless of AOD condition. The main effect of AOD on recognition was non-significant, $F(1, 352) = 0.89$, $p = .346$, $\eta^2p = .002$, indicating that the AI-origin label did not independently elevate advertising recognition. The NAD x AOD interaction on recognition was also non-significant, $F(1, 352) = 0.23$, $p = .632$. Cell means showed a consistent pattern: Control ($M = 2.34$, $SD = 1.12$), AOD-only ($M = 2.51$, $SD = 1.09$), NAD-only ($M = 5.67$, $SD = 0.98$), and Both ($M = 5.81$, $SD = 0.87$). Corresponding analyses for the persuasion-knowledge activation scale replicated these patterns: NAD produced a large significant main effect, $F(1, 352) = 412.56$, $p < .001$, $\eta^2p = .54$ (NAD-present: $M = 5.48$, $SD = 0.89$ vs. NAD-absent: $M = 2.30$, $SD = 0.97$), while neither the AOD main effect, $F(1, 352) = 1.12$, $p = .291$, nor the interaction, $F(1, 352) = 0.41$, $p = .522$, reached significance. H1 was fully supported.

Effects on Perceived Credibility: Main Effects and Interaction

A 2 x 2 ANOVA on the perceived-credibility composite (the mean of ad credibility, source credibility, and perceived-authenticity subscales) yielded significant main effects of NAD, $F(1, 352) = 89.34$, $p < .001$, $\eta^2p = .20$, and AOD, $F(1, 352) = 143.67$, $p < .001$, $\eta^2p = .29$, and a significant NAD x AOD interaction, $F(1, 352) = 8.23$, $p = .004$, $\eta^2p = .023$. Cell means showed the predicted ordering: Control ($M = 5.23$, $SD = 0.87$) > NAD-only ($M = 4.12$, $SD = 0.94$) > AOD-only ($M = 3.89$, $SD = 1.02$) > Both ($M = 2.74$, $SD = 0.91$). The linear trend contrast was significant, $F\_linear(1, 352) = 231.45$, $p < .001$. Simple-effects analysis of the interaction showed that the NAD penalty on credibility was larger in the AOD-present condition ($\Delta = 1.38$) than in the AOD-absent condition ($\Delta = 1.11$), $F(1, 352) = 5.67$, $p = .018$, $\eta^2p = .016$, an over-additive pattern consistent with H4a and inconsistent with H4b. The credibility decrement for the dual-disclosure condition ($\Delta = 2.49$, relative to control) exceeded the arithmetic sum of the two independent decrements ($\Delta\_NAD + \Delta\_AOD = 1.11 + 1.34 = 2.45$), ruling out the transparency-buffer account (H4b). H3 was supported; H4a was supported; H4b was not supported. Analyses of the perceived-authenticity subscale confirmed that AOD produced the largest single effect in the study: $M\_AOD\text{-absent} = 5.38$, $SD = 0.91$ vs. $M\_AOD\text{-present} = 3.00$, $SD = 1.01$, $F(1, 352) = 312.45$, $p < .001$, $\eta^2p = .47$, confirming that the AI label inflicted disproportionate damage on the authenticity dimension of credibility. Table 2 presents the complete 2 x 2 ANOVA results for all dependent variables.

Table 2. Two-Way ANOVA Summary Statistics.

Variable	NAD Effect	AOD Effect	NAD × AOD
Advertising Recognition	$F = 487.23^{***}$ $\eta^2p = .58$	$F = 0.89$, ns	$F = 0.23$, ns
Persuasion Knowledge	$F = 412.56^{***}$ $\eta^2p = .54$	$F = 1.12$, ns	$F = 0.41$, ns
Perceived Authenticity	$F = 3.21$, $p = .074^{a}$	$F = 312.45^{***}$ $\eta^2p = .47$	$F = 1.45$, ns
Perceived Credibility	$F = 89.34^{***}$ $\eta^2p = .20$	$F = 143.67^{***}$ $\eta^2p = .29$	$F = 8.23^{**}$ $\eta^2p = .023$
Perceived Intrusiveness	$F = 67.45^{***}$ $\eta^2p = .16$	$F = 53.78^{***}$ $\eta^2p = .13$	$F = 12.34^{***}$ $\eta^2p = .034$
Feed Enjoyment	$F = 189.34^{***}$ $\eta^2p = .35$	$F = 138.45^{***}$ $\eta^2p = .28$	$F = 15.67^{***}$ $\eta^2p = .043$
Behavioral Intention	$F = 78.23^{***}$ $\eta^2p = .18$	$F = 67.34^{***}$ $\eta^2p = .16$	$F = 2.34$, ns

Mediation Analyses

PROCESS Model 4 with 5,000 bootstrap samples was used to test whether advertising recognition and persuasion knowledge mediated the NAD-to-credibility path. NAD significantly predicted advertising recognition, b = 3.31, SE = 0.15, p < .001, and advertising recognition significantly predicted perceived credibility, b = -0.25, SE = 0.04, p < .001. The indirect effect of NAD on credibility via advertising recognition and persuasion knowledge was significant and negative: b = -0.61, SE = 0.09, 95% CI [-0.82, -0.43]. The direct effect of NAD on credibility, partialling out the mediator, remained significant but attenuated: b = -0.42, SE = 0.12, 95% CI [-0.66, -0.18]. The indirect effect accounted for approximately 59.2% of the total NAD effect on credibility. H2 was supported.

A parallel mediation model (PROCESS Model 4) examined perceived authenticity and persuasion knowledge as simultaneous mediators of the AOD effect on credibility. AOD strongly predicted perceived authenticity, b = -2.38, SE = 0.13, p < .001, which in turn positively predicted credibility, b = 0.43, SE = 0.05, p < .001. The indirect effect of AOD via perceived authenticity was significant: b = -0.82, SE = 0.11, 95% CI [-1.04, -0.60]. The indirect effect via persuasion knowledge was small and non-significant: b = -0.08, SE = 0.06, 95% CI [-0.21, 0.04]. The direct effect of AOD on credibility, controlling for both mediators, was significant: b = -0.31, SE = 0.13, 95% CI [-0.57, -0.06]. The authenticity pathway accounted for 72.6% of the total AOD effect. H5 was supported; the authenticity pathway substantially dominated the persuasion-knowledge pathway for AI-origin disclosure.

Moderating Role of Algorithm Awareness

Algorithm awareness was entered as a continuous moderator in PROCESS Model 58 (moderated mediation). For the NAD pathway, the NAD x Algorithm Awareness product term significantly predicted advertising recognition and persuasion knowledge, b = -0.28, SE = 0.11, t(352) = -2.45, p = .015, indicating that higher algorithm awareness weakened the impact of the commercial label on PK activation. The conditional indirect effect of NAD on credibility via PK was larger for users low in algorithm awareness (M - 1 SD): b = -0.78, SE = 0.14, 95% CI [-1.07, -0.51], and significantly attenuated for users high in algorithm awareness (M + 1 SD): b = -0.43, SE = 0.12, 95% CI [-0.68, -0.20]. The index of moderated mediation was significant: b = -0.18, SE = 0.07, 95% CI [-0.33, -0.04], confirming that algorithm awareness moderated the indirect NAD effect. For the AOD pathway, the AOD x Algorithm Awareness product term was non-significant, b = -0.11, SE = 0.13, t(352) = -0.87, p = .384, and the index of moderated mediation spanned zero: b = -0.08, SE = 0.07, 95% CI [-0.23, 0.07]. H6 was partially supported: algorithm awareness attenuated the persuasion-knowledge costs of commercial disclosure but did not buffer the authenticity-based credibility damage from AI-origin disclosure.

User Experience Outcomes

A 2 x 2 ANOVA on perceived intrusiveness revealed significant main effects of NAD, F(1, 352) = 67.45, p < .001, η^2p = .16 (M_NAD-present = 4.45, SD = 1.21 vs. M_NAD-absent = 2.95, SD = 1.09), and AOD, F(1, 352) = 53.78, p < .001, η^2p = .13 (M_AOD-present = 4.34, SD = 1.18 vs. M_AOD-absent = 3.06, SD = 1.08), as well as a significant interaction, F(1, 352) = 12.34, p < .001, η^2p = .034. Planned contrasts confirmed that the dual-disclosure condition (M = 5.12, SD = 0.97) reported significantly higher intrusiveness than the NAD-only condition (M = 3.78, SD = 0.94, p < .001), the AOD-only condition (M = 3.56, SD = 1.02, p < .001), and the control (M = 2.34, SD = 0.89, p < .001). Feed enjoyment showed the inverse pattern: the dual-disclosure condition reported the lowest enjoyment (M = 2.98, SD = 0.94), significantly below the NAD-only condition (M = 4.12, SD = 0.96, p < .001), the AOD-only condition (M = 4.34, SD = 1.01, p < .001), and the control (M = 5.78, SD = 0.87, p < .001). The overall enjoyment effect was large: F(1, 352) = 189.34, p < .001, η^2p = .350. Attitude toward the ad mirrored the credibility pattern (M_both = 2.61, SD = 0.98 vs. M_control = 5.34, SD = 0.88; M_NAD = 3.95, SD = 0.97; M_AOD = 3.78, SD = 1.04), and behavioral engagement intention was lowest in the dual-disclosure condition (M = 2.48, SD = 1.03). H7 was supported.

Table 3 summarizes the hypotheses, key statistics, and outcomes across all seven tests. Taken together, six of the seven hypotheses (H1, H2, H3, H4a, H5, H7) were supported, H4b was not supported, and H6 received partial support limited to the commercial-disclosure pathway.

Table 3. Summary of Hypothesis Tests and Results.

Hyp.	Finding Tested	Key Statistics	Supported?
H1	NAD → Advertising recognition	F(1, 352) = 487.23, p < .001, η^2p = .58; M_NAD-present = 5.74 vs. M_absent = 2.43	Yes

Hyp.	Finding Tested	Key Statistics	Supported?
H2	PK mediates NAD → Perceived credibility	Indirect b = −0.61, SE = 0.09, 95% CI [−0.82, −0.43]; 59.2% of total effect	Yes
H3	AOD → Perceived credibility (via authenticity)	$F(1, 352) = 143.67$, $p < .001$, $\eta^2p = .29$; Authenticity: $F(1, 352) = 312.45$, $\eta^2p = .47$	Yes
H4a	Dual disclosure: additive credibility cost (M Both = 2.74)	NAD × AOD: $F(1, 352) = 8.23$, $p = .004$, $\eta^2p = .023$; Over-additive: ΔBoth = 2.49 > ΔNAD + ΔAOD = 2.45	Yes
H4b	Transparency buffer: AOD weaker when NAD present	NAD simple effect larger in AOD-present ($\delta = 1.38$ vs. 1.11); $F(1, 352) = 5.67$, $p = .018$: opposite direction	No
H5	Authenticity mediates AOD → Credibility	Indirect b = −0.82, SE = 0.11, 95% CI [−1.04, −0.60]; 72.6% of AOD effect; PK indirect ns	Yes
H6	Algorithm awareness moderates disclosure paths	NAD path index b = −0.18, 95% CI [−0.33, −0.04] (sig.); AOD path index b = −0.08, CI [−0.23, 0.07] (ns)	Partial
H7	Dual disclosure → Highest intrusiveness; lowest enjoyment	Intrusiveness: $F(1, 352) = 12.34$, $p < .001$, $\eta^2p = .034$; Enjoyment: $F(1, 352) = 189.34$, $p < .001$, $\eta^2p = .35$	Yes

Theoretical and Practical Implications

The results extend the CARE model and the Persuasion Knowledge Model into two frontiers that established theory has not jointly addressed: artificial intelligence as the source of covert advertising, and the algorithmic feed as the channel through which both the advertisement and its disclosures are delivered. Where prior disclosure research has treated the commercial label as the principal trigger of persuasion knowledge, the present findings show that an AI-origin label activates a second, functionally distinct persuasion-knowledge pathway operating through perceived authenticity, with an over-additive effect on credibility when both labels are present simultaneously. The attenuation of the commercial-disclosure effect by algorithm awareness, but not the AI-disclosure effect, further clarifies that authenticity-based credibility damage is resistant to the kind of prior priming that buffers recognition-based damage, an important distinction for both theory and practice.

The findings speak directly to a live policy environment. China's Labeling Measures and the EU AI Act compel AI disclosure, but the present results suggest that compliance incurs a substantial credibility cost that is compounded when commercial and AI labels co-occur. For advertisers, the dominance of the authenticity pathway (72.6% of the AOD effect) indicates that humanizing AI-generated creative, or reserving fully synthetic production for functional content less reliant on authenticity, will limit the damage. For platforms and regulators, the large and significant intrusiveness effect in the dual-disclosure condition ($\eta^2p = .034$) provides quantitative grounding for the ergonomic concern that stacked transparency signals disrupt the immersive experience on which short-video engagement depends. And the partial moderation by algorithm awareness underscores the potential value of digital media literacy programs in equipping users to respond to commercial labels more adaptively, even if such literacy cannot substantially buffer the affective impact of AI-origin disclosure.

CONCLUSIONS

Generative AI and mandatory transparency labeling are reshaping the conditions under which persuasion occurs on one of China's most important media platform. AI now allows native advertising to be produced at scale in the organic idiom of the short-video feed, while new disclosure rules require that both its commercial and its synthetic nature be revealed. These forces pull in opposite directions: native advertising profits from covertness, and disclosure exists to dispel it. Reading the convergence through the Persuasion Knowledge Model, the CARE model of advertising recognition, source-credibility theory, and research on algorithm awareness, this paper has argued that an AI-generated native advertisement carrying both a commercial cue and an AI cue activates two distinct strands of persuasion knowledge,

that the resulting credibility cost is concentrated in perceived authenticity, and that the two disclosures are more likely to compound than to cancel, an additive penalty whose magnitude is nonetheless conditioned by content type, by the salience of the label in a fast feed, and by the audience's awareness of the algorithm. The wider lesson is that a transparency regime cannot be evaluated by the candor of its labels alone. On a platform where attention is scarce and authenticity is the currency, whether disclosure informs audiences or merely diminishes the communication it governs depends on how, where, and to whom the labels actually speak.

This paper has reasoned from established theory and from the most directly relevant experimental evidence, but several of its claims are inferential and await direct empirical confirmation in the specific setting of Douyin's feed. Three questions are especially open. The first is the sign of the interaction: whether the additive-penalty default genuinely holds when the two labels are crossed within a realistic vertical feed, or whether a transparency dividend emerges under identifiable conditions is ultimately an empirical matter that the current literature addresses only through extrapolation from single-cue studies conducted on static or Western formats. The second is the direction of the algorithm-awareness moderation, attenuation versus amplification, which the existing scholarship frames but does not resolve. The third concerns habituation over time: as AI labels become ubiquitous under the new mandates, audiences may normalize them, and the cross-sectional logic developed here may not describe a future in which an "AI-generated" tag is as unremarkable as a nutrition label. Cross-cultural comparison between Douyin and TikTok users, who differ in regulatory exposure and platform norms, and designs that combine self-report with behavioral measures of attention and scrolling, would do much to convert the present analysis into tested knowledge.

REFERENCES

Amazeen, M. A., & Wojdynski, B. W. (2020). The effects of disclosure format on native advertising recognition and audience perceptions of legacy and online news publishers. ***Journalism, 21***(12), 1965–1984. https://doi.org/10.1177/1464884918754829

Baek, T. H., Kim, J., & Kim, J. H. (2024). Effect of disclosing AI-generated content on prosocial advertising evaluation. ***International Journal of Advertising, 45***(1), 171–192. https://doi.org/10.1080/02650487.2024.2401319

DeVito, M. A., Birnholtz, J., Hancock, J. T., French, M., & Liu, S. (2018). How people form folk theories of social media feeds and what it means for how we study self-presentation. ***2018 CHI Conference on Human Factors in Computing Systems***. Association for Computing Machinery. Montreal, Canada.

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K., & Sandvig, C. (2015). "I always assumed that I wasn't really that close to [her]": Reasoning about invisible algorithms in news feeds. ***33rd Annual ACM Conference on Human Factors in Computing Systems.*** Association for Computing Machinery. New York, United States.

Flanagin, A. J., & Metzger, M. J. (2000). Perceptions of internet information credibility. ***Journalism & Mass Communication Quarterly, 77***(3), 515–540. https://doi.org/10.1177/107769900007700304

Ford, J., Jain, V., Wadhwani, K., & Gupta, D. G. (2023). AI advertising: An overview and guidelines. ***Journal of Business Research, 166***, Article 114124. https://doi.org/10.1016/j.jbusres.2023.114124

Friestad, M., & Wright, P. (1994). The persuasion knowledge model: How people cope with persuasion attempts. ***Journal of Consumer Research, 21***(1), 1–31. https://doi.org/10.1086/209380

Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). ***Communication and persuasion: Psychological studies of opinion change***. Yale University Press.

Karizat, N., Delmonaco, D., Eslami, M., & Andalibi, N. (2021). Algorithmic folk theories and identity: How TikTok users co-produce knowledge of identity and engage in algorithmic resistance. ***Proceedings of the ACM on Human-Computer Interaction.*** New York, United States.

Lim, S., & Schmälzle, R. (2024). The effect of source disclosure on evaluation of AI-generated messages. ***Computers in Human Behavior: Artificial Humans, 2***(1), Article 100058. https://doi.org/10.1016/j.chbah.2024.100058

Lu, Y., & Pan, J. (2022). The pervasive presence of Chinese government content on Douyin trending videos. ***Computational Communication Research, 4***(1), 68–97. https://doi.org/10.5117/CCR2022.2.002.LU

Taylor, S. H., & Choi, M. (2022). An initial conceptualization of algorithm responsiveness: Comparing perceptions of algorithms across social media platforms. ***Social Media + Society, 8***(4). https://doi.org/10.1177/20563051221144322

Tutaj, K., & van Reijmersdal, E. A. (2012). Effects of online advertising format and persuasion knowledge on audience reactions. ***Journal of Marketing Communications, 18***(1), 5–18. https://doi.org/10.1080/13527266.2011.620765

Voorveld, H. A. M., Meppelink, C. S., & Boerman, S. C. (2024). Consumers' persuasion knowledge of algorithms in social media advertising: Identifying consumer groups based on awareness, appropriateness, and coping ability. ***International Journal of Advertising, 43***(6), 960–986. https://doi.org/10.1080/02650487.2023.2264045

Wang, S., & Huang, G. (2024). The impact of machine authorship on news audience perceptions: A meta-analysis of experimental studies. ***Communication Research, 51***(7), 815–842. https://doi.org/10.1177/00936502241229794

Windels, K., & Porter, L. (2020). Examining consumers' recognition of native and banner advertising on news website home pages. ***Journal of Interactive Advertising, 20***(1), 1–16. https://doi.org/10.1080/15252019.2019.1688737

Wojdynski, B. W., & Evans, N. J. (2020). The covert advertising recognition and effects (CARE) model: Processes of persuasion in native advertising and other masked formats. ***International Journal of Advertising, 39***(1), 4–31. https://doi.org/10.1080/02650487.2019.1658438

Wortel, C., Vanwesenbeeck, I., & Tomas, F. (2024). Made with artificial intelligence: The effect of artificial intelligence disclosures in Instagram advertisements on consumer attitudes. ***Emerging Media, 2***(3), 547–570. https://doi.org/10.1177/27523543241292096

Zarouali, B., Boerman, S. C., & de Vreese, C. H. (2021). Is this recommended by an algorithm? The development and validation of the algorithmic media content awareness scale (AMCA-scale). ***Telematics and Informatics, 62***, Article 101607. https://doi.org/10.1016/j.tele.2021.101607