COMMUNICATION & SOCIETY

Special issue

Algorithmic Truthfulness, Community Notes, and Trust in Greenwashing Claims: Evidence from Aspirational Fashion and Sustainable Luxury Brands

Aitor Gil García
https://orcid.org/0009-0005-1133-7842
aitor.gil@ucjc.edu
Universidad Camilo José Cela

África Presol Herrero
https://orcid.org/0000-0002-7600-8231
apresol@ucjc.edu
Universidad Camilo José Cela

Submitted
December 4, 2025
Approved
April 29, 2026

© 2026
Communication & Society
ISSN 0214-0039
E ISSN 2386-7876
www.communication-society.com

2026 – Vol. 39(1)
pp. 445-463

How to cite this article:
Gil García, A., & Presol Herrero, Á. (2026). Algorithmic truthfulness, community notes, and trust in greenwashing claims: Evidence from aspirational fashion and sustainable luxury brands. *Communication & Society*, *39*(1), 445-463.
https://doi.org/10.15581/003.39.1.030

Abstract

The emergence of aspirational fashion and sustainable luxury brands occurs against the backdrop of growing environmental misinformation and greenwashing in today's communicative ecology. At the same time, various social media platforms (e.g., X) have introduced hybrid verification systems that include Community Notes and combine automated assessments with crowd-sourced judgments. In light of this phenomenon, this paper presents an empirical investigation based on a mixed-method design, which involves (a) a theoretical review, (b) reinterpretation of two existing studies focusing on the use of AI and persuasive communication, and (c) a 2 × 2 factorial experiment (initial n = 240; valid n = 228) that tests the moderating role of Community Notes in the relationship between AI-generated messages and their credibility. According to the findings, algorithmic personalization increases persuasiveness when AI-generated messages are supported by credible information; otherwise, they heighten concerns about greenwashing. Community Notes are interpreted as epistemic signals: they increase trust when a message is authentic and trigger more critical judgment when it contradicts expectations. Based on the observations, this research contributes to the scholarly literature by introducing an AI + Community framework that conceptualizes how AI, human judgment, and crowd wisdom interact to form sustainable credibility. Possible implications of the study center on the criteria for efficient communication governance in the age of automation.

Keywords

algorithmic veracity; greenwashing; Community Notes; credibility; sustainability; artificial intelligence; collaborative verification.

1. Introduction

Over the past decade, sustainability has become a central dimension of corporate discourse across sectors such as fashion and aspirational luxury, driven by consumers who are increasingly aware of the environmental consequences of their purchasing decisions (Universitat Oberta de Catalunya, 2023). In this context, brands intensively communicate their

"green" credentials to reinforce both market positioning and commercial performance. However, this shift has been accompanied by a significant rise in greenwashing practices, understood as strategies that exaggerate or distort environmental benefits (Delmas & Burbano, 2011). Recent studies confirm the expansion of such practices in the industry (Koch & Denner, 2025; Santos et al., 2024), contributing to new forms of ecological misinformation and eroding public trust. In Spain, more than 80% of citizens perceive corporate sustainability as "just marketing," lacking verifiable support (Universitat Oberta de Catalunya, 2023), reinforcing the idea that the absence of traceability undermines brand legitimacy (Jian & Zhong, 2025).

At the same time, digital platforms are transforming their moderation systems in response to the proliferation of misinformation (Matassi & Boczkowski, 2020). Traditionally, verification relied on professional fact-checkers[1] (Graves, 2016). Still, the volume and speed of content have driven the emergence of automated verification systems based on artificial intelligence, capable of scaling these processes in massive digital environments (Mittelstadt et al., 2016; Floridi et al., 2018).

These use artificial intelligence to detect misleading claims and compare them with verified databases (Mittelstadt et al., 2016; Floridi et al., 2018). However, they present significant limitations, such as difficulty interpreting complex contexts or reproducing biases (Noble, 2018; Eubanks, 2018).

Given these limitations, hybrid models have emerged that combine automation, human supervision, and user participation (Sendra-Duro, 2025). Community Notes, on X (formerly Twitter), is the most representative example. Originating from the Birdwatch program (2021), it allows users to add contextual notes to potentially misleading posts.

These notes are only displayed when there is consensus among participants with different perspectives, incorporating mechanisms of "collective wisdom" in verification. Thus, Community Notes does not replace traditional fact-checking, but rather reconfigures it into a collaborative, algorithmically assisted model (La Barbera et al., 2024; Gao et al., 2024). Evidence suggests that this approach can improve perceptions of impartiality, though it also poses challenges regarding epistemic authority and verification quality (Drolsbach et al., 2024).

This article is situated at the intersection of two transformations: the dispute over the truthfulness of sustainable communication and the emergence of hybrid verification models. In this context, the following research question is posed: *To what extent can mechanisms like Community Notes modify perceptions of the truthfulness, transparency, and legitimacy of environmental messages in brands' sustainable fashion aspirations?* The issue is particularly relevant in environments where users show skepticism toward automation (Longoni et al., 2019) and face highly polarized information scenarios (Zarouali et al., 2022).

Despite the growing volume of studies on artificial intelligence in marketing, sustainability, and information verification, these strands have evolved in parallel and with little integration. The AI literature highlights its ability to enhance personalization and persuasive effectiveness (Davenport et al., 2020; Kumar et al., 2021), while sustainability studies show that greenwashing negatively impacts consumer trust (Delmas & Burbano, 2011; Santos et al., 2024).

1 Fact-checking is defined as the systematic process of verifying the factual accuracy of public statements or disseminated content, based on comparison with verifiable sources, whose objective is to assess their truthfulness and contribute to correcting misinformation in media and digital environments (Graves, 2019).

ISSN 2386-7876 – © 2026 Communication & Society, 39(1), 445-463

Meanwhile, verification research has analyzed both automated systems and professional fact-checking (Graves, 2016; Lewandowsky et al., 2020). However, there is limited empirical evidence on how these dynamics interact in digital environments where automation and collaborative verification converge.

This gap is especially relevant in the sustainable fashion sector, where credibility depends not only on the message's content but also on the informational ecosystem in which it is disseminated. In this sense, the present study proposes an integrative approach that combines persuasive communication, algorithmic veracity, and hybrid moderation models (**Sundar**, 2020; Afroogh et al., 2024; Sebastião & Dias, 2025).

Methodologically, the research articulates the reinterpretation of two previous studies on artificial intelligence and persuasive communication with an original factorial experiment, which allows for jointly analyzing the roles of algorithmic personalization, information traceability, and community verification in credibility construction.

Finally, from an applied perspective, the study offers implications relevant for different stakeholders. For brands, it offers guidelines to communicate sustainability with greater coherence and transparency. For platforms, it enables evaluating the potential of hybrid verification models. For consumers and regulators, it provides a critical basis for understanding the effectiveness of these mechanisms in combating greenwashing and environmental misinformation.

2. The theoretical framework

The theoretical framework integrates three key dimensions to understand the credibility of environmental messages in automated digital environments: artificial intelligence in persuasive personalization, epistemic challenges related to sustainability and greenwashing, and the consolidation of hybrid verification models that align algorithmic and community participation. These perspectives situate the phenomenon in a context where truthfulness, traceability, and communicative legitimacy are increasingly contested.

2.1. *Artificial intelligence in persuasive communication*

The incorporation of artificial intelligence in brand communication has reconfigured persuasive dynamics, enabling advanced levels of algorithmic personalization. These systems facilitate micro-audience segmentation, message adaptation to user profiles, and content optimization through predictive models (Davenport et al., 2020; Kumar et al., 2021). Although these capabilities increase persuasive effectiveness, they also introduce risks, including opacity and the potential amplification of biases when truthfulness is not integrated into the campaign design (Zarouali et al., 2022).

Previous research has evidenced a structural tension: while professionals value AI for its ability to increase relevance and segmentation, they also warn that it can intensify ethical errors when 'green' messages are optimized for clicks without guaranteeing their accuracy. From the consumer perspective, there is a penalty for communications perceived as incoherent or overly automated, in line with resistance to decisions attributed to non-human agents (Longoni et al., 2019).

This ambivalence is reflected in Figure 1-A1 of Appendix 1, which synthesizes the main results reported by Gil García and Presol Herrero (2026). Derived from a mixed-methods study involving interviews with 50 marketing managers and a survey of 500 consumers, the figure shows simultaneous increases in trust in AI (from 40.3% to 65.2%) and satisfaction with personalization (from 48.3% to 74.6%), alongside an increase in privacy concerns (from 22.2% to 30.5%). Taken together, these results suggest that the adoption of

ISSN 2386-7876 – © 2026 Communication & Society, 39(1), 445-463

automation is conditional: users value its benefits but also demand transparency, traceability, and human oversight.

2.2. *Sustainability, truth, and greenwashing in aspirational brands*

Environmental misinformation takes many forms, but greenwashing is its most widespread expression in brand communication. This phenomenon encompasses practices that appear to be sustainable initiatives but lack sufficient backing, supported by vague or misleading assertions (Delmas & Burbano, 2011; Santos et al., 2024). This dynamic generates tensions between discourse and reality, compromising the communicative legitimacy of companies (Fernández-Fernández & Pinillos, 2011).

The magnitude of the problem has been widely documented: the European Commission estimates that 42% of analyzed environmental claims are misleading or exaggerated, constituting a form of epistemic injustice (Fricker, 2007). This context affects both consumers, who make uninformed decisions, and sustainable companies, facing widespread skepticism (Jian & Zhong, 2025). In sectors such as fashion, the heavy use of green claims amplifies perceptions of manipulation and erodes trust (Santos et al., 2024).

Moreover, consumer environmental knowledge increases their ability to detect inconsistencies, reducing purchase intent and damaging brand reputation (Koch & Denner, 2025; Soundarya, 2025). In response, stricter regulatory frameworks have emerged that require verifiability and traceability of environmental claims (European Union, 2024). However, information asymmetries and competitive pressure dynamics continue to favor overselling[2] practices, even in companies with legitimate intentions (Couldry & Mejías, 2019; Zuboff, 2019).

This scenario underscores the need for mechanisms to verify and contextualize information in real time, especially on social media, where brand messages coexist with interpretations and corrections from users and automated systems.

2.3. *Hybrid moderation, community verification, and Community Notes*

In this context, automated fact-checking has evolved from traditional verification models, incorporating AI-based systems capable of detecting potentially misleading statements, cross-checking them against verified sources, and identifying disinformation patterns at scale (Floridi et al., 2018; Mittelstadt et al., 2016). This process has given rise to the concept of "algorithmic veracity," understood as the set of procedures through which automated systems prioritize and contextualize information in digital environments. Nevertheless, automation presents limitations, especially in contextual interpretation and in reproducing biases (Noble, 2018; Eubanks, 2018). These limitations have favored the emergence of hybrid models that combine algorithmic efficiency with collective intelligence.

In this context, Community Notes on X serves as a representative example, as it allows users to add contextual notes to potentially misleading posts. These notes are displayed only when consensus is reached among participants with different perspectives, integrating mechanisms of "collective wisdom" into the verification process. The system thus combines algorithmic content identification with community evaluation, generating a distributed verification model.

Empirical evidence suggests that notes with argued context and verifiable sources are perceived as more reliable than generic warnings (Drolsbach et al., 2024; Pilarski et al., 2023)

2 Overselling refers to the practice of exaggerating or overstating the features or benefits of a product or message beyond what is verifiable, to increase its persuasive appeal or competitive advantage (Lyon & Montgomery, 2015; Delmas & Burbano, 2011).

ISSN 2386-7876 – © 2026 Communication & Society, 39(1), 445-463

and that their presence increases the likelihood of content correction (Gao et al., 2024). However, tensions persist around epistemic authority, as some members of the public continue to trust professional verifiers more than collaborative systems (Jia & Lee, 2024).

From media literacy, it has been shown that corrective explanations reduce the persistence of erroneous beliefs (Lewandowsky et al., 2020), which is especially relevant in the field of sustainability. In this sense, hybrid models introduce key implications for brands: they increase the reputational cost of greenwashing, generate new layers of public interpretation, and establish emerging verification standards that directly influence perceived credibility.

Together, these systems provide an appropriate framework for analyzing how the interaction among artificial intelligence, community, and sustainability shapes the veracity and legitimacy of messages in digital environments.

3. Research questions, objectives, and hypotheses

The intersection of artificial intelligence, sustainability, and collaborative verification raises relevant questions about how credibility is currently built in the environmental communications of aspirational brands.

Based on the theoretical framework, this research formulates the following questions, objectives, and conceptual hypotheses.

3.1. *Research questions*

RQ1. How do artificial intelligence systems applied to the persuasive communication influence the perception of truthfulness and credibility of environmental messages issued by aspirational sustainable fashion and luxury brands?

RQ2. In what ways do greenwashing and the lack of information traceability erode the epistemic legitimacy of these brands in digital environments?

RQ3. To what extent do hybrid verification models, particularly Community Notes, contribute to strengthening, correcting, or redistributing epistemic authority in the public evaluation of sustainability claims?

RQ4. How do algorithmic personalization, verification, community verification, and consumer sensitivity shape credibility and the intention to support aspirational sustainable brands?

3.2. *Research objectives*

SO1. Analyze the role of artificial intelligence in shaping the credibility, transparency, and interpretation of environmental messages on social platforms.

SO2. Examine how greenwashing affects perceived truthfulness, trust, and the legitimacy of aspirational sustainable fashion and luxury brands.

SO3. Evaluate the potential f hybrid moderation models, in particular Community Notes as mechanisms of correction, contextualization, and redistribution of epistemic authority in the sustainability field.

SO4. Integrate conceptual and empirical findings from previous research to propose an interpretive model that explains how AI + community influence credibility and epistemic legitimacy of "green" communications.

3.3. *Hypotheses*

H1. Algorithmic personalization of environmental messages increases the perceived relevance, but it can also raise suspicion when not accompanied by traceability and verifiable evidence.

ISSN 2386-7876 – © 2026 Communication & Society, 39(1), 445-463

H2. The detection of inconsistencies between the message and practice, especially when amplified by community interventions, reduces perceived credibility and weakens brands' epistemic legitimacy.

H3. Hybrid verification systems, such as Community Notes, generate truth cues (cognitive signals that guide the assessment of message veracity) that can improve public evaluation of sustainable messages, both to identify greenwashing and to reinforce authentic communications.

H4. The credibility assigned to an environmental message is significantly influenced by the interaction between factors such as the algorithmic transparency of the communicative process, the quality and neutrality of community verification, and the user's prior sensitivity toward sustainability, without excluding the possible influence of other contextual and cognitive factors not analyzed in this study.

4. Methodology

This study adopts a hybrid methodological approach based on theoretical-empirical triangulation, aimed at analyzing how the interaction among artificial intelligence, greenwashing, and collaborative verification shapes the credibility of environmental messages in aspirational and sustainable luxury brands.

4.1. *General methodological design*

The methodological design follows a sequential exploratory mixed-methods approach (DEXPLOR-S), suitable for analyzing complex socio-technical phenomena that require an initial phase of conceptual and empirical exploration, followed by a phase of validation and testing (Bryman, 2016; Creswell & Plano Clark, 2017). The research articulates three levels of evidence: theoretical review, reinterpretation of previous studies, and experimentation. This integration enables the formulation of a hybrid interpretive model (AI + Community) to understand how veracity is constructed in automated digital environments.

The methodological strategy is structured in three interconnected phases. First, a theoretical review of five areas was conducted: artificial intelligence applied to persuasive communication; credibility and trust in algorithmic systems; corporate sustainability and greenwashing; epistemic justice and informational legitimacy; and hybrid models of collaborative verification, with emphasis on Community Notes. This pace allowed identifying key constructs and building the analytical framework of the study (Mittelstadt et al., 2016; Afroogh et al., 2024; Miceli & Posada, 2022).

Secondly, the analytical reinterpretation of two previous investigations, which focused respectively on the impact of AI on the persuasive communication of startups and SMEs, and on the construction of credibility, authenticity, and transparency in aspirational sustainable fashion brands. This rereading allowed for the identification of empirically grounded concepts: algorithmic personalization, informational traceability, algorithmic truthfulness, and distributed epistemic governance. This served as a basis for formulating hypotheses and the experimental design.

Thirdly, a controlled, exploratory-explanatory experiment with a 2×2 factorial design was developed to contrast the effect of community verification on the perceived credibility of environmental messages generated or amplified by AI. Two variables were manipulated: the truthfulness of the environmental claim (authentic vs. greenwashing) and the presence or absence of visible Community Notes claims. 240 social media users residing in Spain (228 valid cases) participated, randomly assigned to one of the four experimental conditions. The stimuli

ISSN 2386-7876 – © 2026 Communication & Society, 39(1), 445-463

were simulated posts with fictitious brands, and, in intervention conditions, the community note included verifiable references and a neutral tone.

4.2. Study 1: Persuasive communication with AI in digital environments

The first reinterpretation study uses a sequential exploratory mixed design (DEXPLOR-S) to analyze the impact of AI on persuasive communication in startups and small businesses. The qualitative phase included 50 semi-structured interviews with marketing and communications managers, while the quantitative phase consisted of a survey applied to 500 consumers (Gil & Presol, 2026). Variables measured included perceived personalization, trust, traceability, informational clarity, and persuasive impact (Creswell & Plano Clark, 2017).

Figure 2-A1 of Appendix 1 synthesizes the variation in three key indicators of performance and customer experience: conversion rate, engagement, and satisfaction before and after the implementation of AI-based communication strategies, based on results reported by Gil and Presol (2026). The three indicators show increases after adopting these technologies: conversion rate rises from 10.2% to 30.5%, engagement from 15.4% to 40.3%, and customer satisfaction from 70.1% to 90.7%.

Overall, the findings revealed a structural ambivalence: automation improves segmentation, conversion, and satisfaction, but it also raises concerns about opacity and narrative ethics. This tension was central to grounding the hypotheses tested in Study 3. Details of the fieldwork, together with the profiles of the interview participants in Study 1, are presented in Tables 1-A1, 2-A1, and 3-A1 of Appendix 1.

4.3. *Study 2: AI, truth, and sustainability in aspirational brands*

The second study analyzes how AI reconfigures perceptions of authenticity and credibility among Spanish sustainable fashion brands through a sequential exploratory mixed-methods design. The qualitative phase consisted of 10 semi-structured interviews with sector executives, which were coded thematically, and the quantitative phase involved a survey of 500 consumers, of whom 274 valid cases were retained. Participant profiles, sampling criteria, and fieldwork details for Study 2 are reported in Tables 1-A2, 2-A2, and 3-A2 of Appendix 2.

The scales used assessed algorithmic transparency, trust, message relevance, perception of greenwashing, privacy sensitivity, and persuasiveness. Validated instruments and analyses with Jamovi 2.4 were used, including t-tests to compare pre-post means, Cronbach's alpha coefficients, and calculations of effect size (Cohen's *d*), following the recommendations of Hair et al. (2019).

Table 3 – Appendix 2 summarizes the main changes perceived after interacting with AI-assisted messages. Significant increases were observed in all analyzed variables ($p < .001$), with particular increases in perceived relevance (from 3.21 to 3.76; $d = .71$) and in self-reported engagement (from 2.98 to 3.58; $d = .75$), as well as improvements in persuasive effectiveness (from 3.05 to 3.52; $d = .56$) and brand satisfaction (from 3.34 to 3.72; $d = .49$).

These results indicate that algorithmic personalization enhances the effectiveness of communication and user experience, though qualitative evidence suggests these benefits only consolidate when accompanied by verifiable traceability and human control. In this sense, credibility does not depend solely on the message's content but also on the epistemic architecture that structures the relationship among technology, brand, and audience.

4.4. *Study 3: Experiment on Community Notes and Sustainable Credibility*

The third study expands on the findings of the previous phases through an experiment aimed at analyzing to what extent a community verification intervention, in this case Community

ISSN 2386-7876 – © 2026 Communication & Society, 39(1), 445-463

Notes in X, can modify the perceived credibility of AI-generated environmental messages. To this end, an online experiment with a 2×2 factorial design was conducted, in which the truthfulness of the environmental claim (authentic vs. greenwashing) and the presence or absence of a community note with verifiable information were manipulated.

The initial sample consisted of 240 participants residing in Spain, recruited through an online panel with age and gender quotas designed to approximate the adult digital population. Assignment to the four experimental conditions was randomized using the Qualtrics distribution system. After data cleaning, the final valid sample consisted of 228 participants. Most were regular social media users, with a mean age of 30-40 years, a balanced gender distribution, and predominantly secondary or university-level education. This profile was considered appropriate for evaluating sustainability-related messages in digital environments. The sample size is consistent with similar experimental designs aimed at detecting moderate effect sizes (Hair et al., 2019). To ensure response quality, attention filters and a minimum stimulus-exposure time were implemented, and 12 cases were excluded for failing to meet response-quality criteria (i.e., inconsistent responses or insufficient response times).

The experiment was implemented entirely in Qualtrics and lasted approximately between six and eight minutes. Each participant viewed a single simulated post, created from real communication patterns in a sustainable fashion, but adapted with fictitious brands to avoid biases associated with familiarity or prior reputation. The stimuli replicated typical social media posts and included product image, logo, brief copy, and environmental claim. In the verification condition, the content included a Community Notes note that added context and external references.

Figure 1-A3 of Appendix 3 presents representative examples of the stimuli used, with the aim of reinforcing methodological transparency and visualizing the differences between experimental conditions (authentic claim vs. greenwashing; with vs. without Community Notes). The stimuli shown are simulations designed specifically for the experiment, based on real communication patterns in social networks, and keeping constant the key elements (presence/absence of verification through Community Notes).

The creatives maintained a consistent aspirational aesthetic and incorporated common sector claims, such as “organic fabric without harmful chemicals” or “produced with 100% renewable energy.” In the verification conditions, the Community Notes appeared to be written in a neutral tone and were accompanied by verifiable references from external audits, public databases, or digital product passports. Before the main study, the stimuli underwent a pre-test with 32 participants, which confirmed their realism, clarity, and visual coherence.

After viewing the assigned stimulus, participants completed a questionnaire using a 7-point Likert scale. The measures assessed perceived credibility (Choi et al., 2025), trust in the message, informational clarity, perceived greenwashing (Lloyd et al., 2025), and purchase intention, the latter treated as an exploratory variable. Sociodemographic variables and control variables related to familiarity with generative AI and attitudes toward sustainability were also collected.

To validate participants’ correct perception of the experimental manipulations and to reinforce the design’s internal validity, manipulation checks were incorporated to verify that participants distinguished between authentic messages and greenwashing messages, as well as between the presence and absence of the Community Notes intervention.

The study was conceived as a controlled, exploratory-explanatory experiment aimed at identifying preliminary trends rather than making population-level inferences. Its purpose

ISSN 2386-7876 – © 2026 Communication & Society, 39(1), 445-463

was to evaluate whether a community intervention can operate as a corrective mechanism in contexts dominated by algorithmic automation, in line with emerging research on distributed epistemic governance and hybrid verification (Gao et al., 2024; Drolsbach et al., 2024). Taken together, the experiment enables analysis of how the combination of generative AI and collaborative verification shapes the construction of credibility in environmental messages.

5. Results

The 2 × 2 factorial experiment enabled evaluation of the effects of claim truthfulness and collaborative verification via Community Notes across five evaluative dimensions. Analyses were conducted in Jamovi 2.4. Prior to hypothesis testing, the assumptions of residual normality and homogeneity of variances were assessed through visual inspection and standard diagnostic tests, with no substantial violations detected that would compromise interpretation. The full set of assumption statistics is reported in the supplementary material.

5.1. *Reliability of the scales and manipulation checks*

Internal consistency of the scales, each composed of three Likert-type items, was assessed using Cronbach's alpha. All variables showed adequate reliability: credibility (α = .89), trust (α = .87), informational clarity (α = .84), perceived greenwashing (α = .86), and purchase intention (α = .88). All coefficients exceeded the recommended threshold of .70, confirming the reliability of the measures used. *Table 1* summarizes these results.

Table 1. Reliability of the scales used in the study

Variable	No. of items	Cronbach's α
Credibility	3	.89
Trust	3	.87
Informational clarity	3	.84
Perception of greenwashing	3	.86
Purchase intention	3	.88

Source: Own elaboration from the experiment data (Study 3).

Additionally, the manipulation checks confirmed that participants correctly distinguished between the experimental conditions. Significant differences were observed in perceived message veracity between the authentic-claim condition (M = 5.52, SD = 0.88) and the greenwashing condition (M = 3.71, SD = 1.02), $F(1, 224) = 112.43$, $p < .001$, $\eta^2_p = .33$. Similarly, the presence of community verification was correctly identified, with higher scores in the Community Notes conditions (M = 5.34, SD = 0.91) than in the no-note conditions (M = 2.11, SD = 1.04), $F(1, 224) = 198.76$, $p < .001$, $\eta^2_p = .47$. These results support the successful implementation of the manipulations and, consequently, the internal validity of the experimental design.

5.2. *Main results of the 2 × 2 factorial ANOVA*

Table 2 presents the results of the 2 × 2 factorial ANOVA for the main dependent variables. Overall, significant main effects of message veracity were observed across all variables analyzed, along with smaller effects associated with the presence of Community Notes and the interaction between the two factors.

ISSN 2386-7876 – © 2026 Communication & Society, 39(1), 445-463

Table 2. Results of the 2×2 factorial ANOVA for the main dependent variables (n = 228)

Variable	Factor	df	MS	F	p	η^2_p
Credibility	Truthfulness	1, 224	42.18	58.34	< .001	.21
	Community Notes	1, 224	6.44	8.92	.003	.04
	Interaction	1, 224	7.38	10.21	.002	.04
Confidence	Truthfulness	1, 224	39.02	54.11	< .001	.19
	Community Notes	1, 224	5.65	7.84	.006	.03
	Interaction	1, 224	6.12	8.47	.004	.04
Informative clarity	Truthfulness	1, 224	34.77	49.28	< .001	.18
	Community Notes	1, 224	2.85	4.03	.045	.02
	Interaction	1, 224	3.94	5.58	.019	.02
Perception of greenwashing	Truthfulness	1, 224	30.15	43.07	< .001	0.16
	Community Notes	1, 224	4.21	6.02	.015	.03
	Interaction	1, 224	5.03	7.18	.008	.03
Intention of purchase	Truthfulness	1, 224	28.63	41.12	< .001	.16
	Community Notes	1, 224	3.12	4.48	.035	.02
	Interaction	1, 224	8.21	11.76	< .001	.05

Note. df = degrees of freedom; MS = mean square; F = ANOVA F statistic; p = significance level; η^2_p = partial eta squared. Reference values for η^2_p are .01 (small), .06 (medium), and .14 (large) (Cohen, 1988).

Source: Own elaboration from the experiment data (Study 3).

In order to facilitate the interpretation of the main effects and interactions, *Table 3* presents the means and standard deviations of the dependent variables in each experimental condition. These values allow direct observation of the underlying patterns in the ANOVA results, especially regarding the differential effect of community verification depending on the truthfulness of the message.

Table 3. Means and standard deviations of the dependent variables by experimental condition (n = 228)

Variable	Authentic without note	Authentic with note	Greenwashing without note	Greenwashing with note
Credibility	M = 5.82 (SD = 0.91)	M = 6.18 (SD = 0.84)	M = 3.94 (SD = 1.02)	M = 3.21 (SD = 1.08)
Trust	M = 5.76 (SD = 0.95)	M = 6.10 (SD = 0.88)	M = 3.88 (SD = 1.05)	M = 3.15 (SD = 1.10)
Informational clarity	M = 5.68 (SD = 0.93)	M = 5.95 (SD = 0.89)	M = 4.02 (SD = 1.01)	M = 3.55 (SD = 1.06)
Perceived greenwashing	M = 2.18 (SD = 0.88)	M = 1.82 (SD = 0.80)	M = 4.92 (SD = 1.00)	M = 5.48 (SD = 0.95)
Purchase intention	M = 5.54 (SD = 0.97)	M = 5.98 (SD = 0.90)	M = 3.72 (SD = 1.08)	M = 3.05 (SD = 1.12)

Note: M = mean; SD = standard deviation. The experimental conditions correspond to the combination of claim veracity (authentic vs. greenwashing) and the presence of Community Notes (with vs. without note).

Source: Own elaboration based on experimental data (Study 3).

5.3. Main effects: content and context both matter

The analysis of variance revealed significant main effects of message veracity across all dependent variables. As shown in *Table 3*, conditions with authentic claims consistently yielded higher scores in perceived credibility (M = 5.82–6.18), trust (M = 5.76–6.10), informational clarity (M = 5.68–5.95), and purchase intention (M = 5.54–5.98) compared to greenwashing conditions (credibility: M = 3.21–3.94; trust: M = 3.15–3.88; informational clarity: M = 3.55–4.02; purchase intention: M = 3.05–3.72). Conversely, perceived

ISSN 2386-7876 – © 2026 Communication & Society, 39(1), 445-463

greenwashing was lower in authentic conditions (M = 1.82–2.18) and higher in greenwashing conditions (M = 4.92–5.48).

These differences were statistically significant in all cases: credibility, $F(1, 224) = 58.34$, $p < .001$, $\eta^2_p = .21$; trust, $F(1, 224) = 54.11$, $p < .001$, $\eta^2_p = .19$; informational clarity, $F(1, 224) = 49.28$, $p < .001$, $\eta^2_p = .18$; perceived greenwashing, $F(1, 224) = 43.07$, $p < .001$, $\eta^2_p = .16$; and purchase intention, $F(1, 224) = 41.12$, $p < .001$, $\eta^2_p = .16$. Overall, the effect sizes for message veracity ranged from .16 to .21, indicating a substantial impact of message content on stimulus evaluation.

Significant main effects of Community Notes were also observed; however, these effects should be interpreted in light of the significant interactions with message veracity, discussed in the following section. Specifically, Community Notes affected credibility, $F(1, 224) = 8.92$, $p = .003$, $\eta^2_p = .04$; trust, $F(1, 224) = 7.84$, $p = .006$, $\eta^2_p = .03$; informational clarity, $F(1, 224) = 4.03$, $p = .045$, $\eta^2_p = .02$; perceived greenwashing, $F(1, 224) = 6.02$, $p = .015$, $\eta^2_p = .03$; and purchase intention, $F(1, 224) = 4.48$, $p = .035$, $\eta^2_p = .02$. Overall, these results suggest that Community Notes function as relevant interpretive cues in message evaluation, although their effect is not uniform and depends on the type of claim presented.

5.4. *Interaction effects: verification as amplifier or corrector*

The interaction between message veracity and the presence of Community Notes was significant across all five dependent variables, indicating that the effect of the note varies depending on whether the claim is authentic or represents greenwashing.

For credibility, a significant interaction was observed, $F(1, 224) = 10.21$, $p = .002$, $\eta^2_p = .04$. As shown in *Table 3*, when the claim was authentic, the presence of the note increased credibility (M without note = 5.82, M with note = 6.18), acting as an epistemic amplifier. In contrast, under greenwashing conditions, the note reduced perceived credibility (M without note = 3.94, M with note = 3.21), functioning as a corrective mechanism.

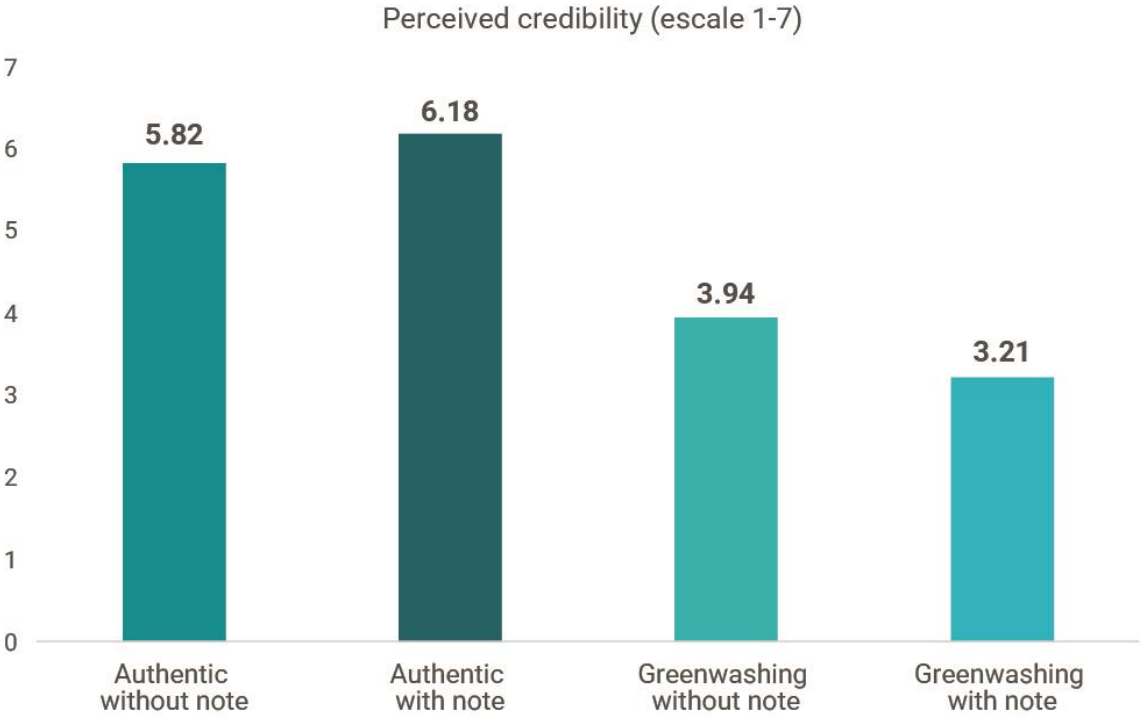
This same pattern was replicated across the other variables. For trust, the interaction was significant, $F(1, 224) = 8.47$, $p = .004$, $\eta^2_p = .04$: the note increased trust in authentic messages (M without note = 5.76, M with note = 6.10) but decreased it in greenwashing messages (M without note = 3.88, M with note = 3.15). For informational clarity, the interaction was also significant, $F(1, 224) = 5.58$, $p = .019$, $\eta^2_p = .02$: Community Notes improved clarity in authentic messages (M without note = 5.68, M with note = 5.95) while reducing it in inconsistent messages (M without note = 4.02, M with note = 3.55). For purchase intention, the interaction was again significant, $F(1, 224) = 11.76$, $p < .001$, $\eta^2_p = .05$: evaluations increased under authentic conditions (M without note = 5.54, M with note = 5.98) and decreased under greenwashing conditions (M without note = 3.72, M with note = 3.05).

Similarly, for perceived greenwashing, a significant interaction was observed, $F(1, 224) = 7.18$, $p = .008$, $\eta^2_p = .03$. In authentic messages, the note reduced perceived greenwashing (M without note = 2.18, M with note = 1.82), whereas in inconsistent messages it increased it (M without note = 4.92, M with note = 5.48), intensifying the critical evaluation of the content.

As illustrated in *Figure 1*, credibility reaches its highest value in the authentic message condition with Community Notes and its lowest value in the greenwashing condition with notes. Overall, this pattern confirms that community-based verification does not operate uniformly but instead modulates message interpretation as a function of veracity—reinforcing credibility when the content is consistent and amplifying critical responses when it is incoherent. These findings support the logic proposed in H3 and H4.

ISSN 2386-7876 – © 2026 Communication & Society, 39(1), 445-463

Figure 1. Interaction between message veracity and the presence of Community Notes in perceived credibility (n = 228)



Source: Own elaboration based on the experiment data (Study 3).

6. Discussion

The experiment results confirm that the credibility of environmental messages generated or amplified through artificial intelligence is not an inherent attribute of the content, but a relational process conditioned by traceability, verification signals, and the interpretive context.

In line with H2, H3 and H4, the truthfulness of the claim and the presence of Community Notes act as key factors in building trust, evidencing that the evaluation of messages depends on both the content and the epistemic architecture in which they are embedded.

6.1. *Validation of patterns identified in previous studies*

The experiment reinforces the findings of Studies 1 and 2. Whereas Study 1 showed that algorithmic personalization increases persuasive effectiveness but raises concerns in the absence of traceability, and Study 2 demonstrated the fragility of credibility in opaque contexts, the experiment confirms that community verification meaningfully modulates this relationship.

More specifically, Community Notes operate with a dual effect: they amplify credibility in truthful messages while reducing it in greenwashing contexts. This pattern confirms their role as truth cues, that is, signals that orient user interpretation toward trust or suspicion, thereby providing empirical support for the mediating role proposed in H3.

6.2. *Redistribution of epistemic authority in digital environments*

One of the main contributions of the study is the evidence that collaborative verification systems redistribute epistemic authority in digital environments. Community Notes do not

ISSN 2386-7876 – © 2026 Communication & Society, 39(1), 445-463

replace brand communication but introduce an additional layer of validation through user–algorithmic system interactions.

The results show that these interventions reinforce trust when the content is consistent and activate critical evaluation when there are inconsistencies, functioning as mechanisms of distributed governance of information. This finding is consistent with recent research on hybrid verification and community participation (Gao et al., 2024; Miceli & Posada, 2022).

6.3. *The role of transparency as a key moderator*

The study confirms that transparency acts as a central moderator in the relationship between automation and credibility (H4). The presence of verifiable signals, such as neutral language, accessible references, or verifiable sources, favors more reflective evaluations and reduces user uncertainty.

On the contrary, the absence of these signals increases suspicion and limits the persuasive effectiveness, even in messages seeking to communicate sustainability. This suggests that credibility in automated environments depends less on technological sophistication than on the ability to generate visible traceability in communicative processes (Mittelstadt et al., 2016).

6.4. *Theoretical and applied implications*

From a theoretical perspective, the results contribute to expanding traditional credibility models by incorporating algorithmic and participatory dimensions into their conceptualization. The evidence supports the need to understand credibility as a hybrid phenomenon in which not only the source and the message, but also verification systems and community interaction, shape evaluative judgments.

In the applied plane, the findings are especially relevant for sectors exposed to the risk of greenwashing, such as sustainable fashion. Community Notes not only improve message understanding but also act as correction and reinforcement mechanisms for coherent communicative practices.

Overall, the results reinforce the potential of the AI + Community model as a way to strengthen credibility and communicative legitimacy in digital environments characterized by automation.

7. Conclusions and the AI + Community conceptual model

This section synthesizes the study's principal contributions by integrating its objectives, empirical findings, and tested hypotheses. On the basis of this triangulated evidence, the research questions are addressed, and the AI + Community conceptual model is advanced as an interpretive framework for understanding how credibility is constructed around sustainable messages in automated digital environments.

7.1. *Relation between objectives, empirical findings, and hypotheses*

Table 4 summarizes the correspondence between the specific objectives, the results obtained in the three phases of the study (qualitative, quantitative, and experimental) and the formulated hypotheses. This integration demonstrates the coherence of the methodological design and confirms that the hypotheses are supported by convergent evidence from different approaches.

ISSN 2386-7876 – © 2026 Communication & Society, 39(1), 445-463

Table 4. Correspondence between specific objectives, empirical findings, and tested hypotheses

Specific Objective	Empirical Findings	Associated Hypothesis	Hypothesis Testing Results
SO1. Analyze the role of artificial intelligence in the credibility and transparency of environmental messages.	AI increases communicative effectiveness when accompanied by traceability; without it, it triggers skepticism.	H1. Algorithmic personalization of environmental messages increases perceived relevance, but may also raise suspicion when not supported by traceability and verifiable evidence.	Confirmed. Algorithmic personalization significantly increased perceived relevance (d = 0.71) and persuasive effectiveness (d = 0.56). However, its effect depends on the presence of traceability and verification, as their absence increases perceived artificiality and reduces credibility.
SO2. Examine how greenwashing affects perceived veracity and the legitimacy of sustainable brands.	Greenwashing without traceability significantly reduces trust and perceived legitimacy.	H2. The detection of inconsistencies between message and practice, especially when amplified by community interventions, reduces perceived credibility and weakens the epistemic legitimacy of brands.	Confirmed. Greenwashing messages showed significantly lower values for credibility, trust, informational clarity, and purchase intention than authentic messages ($p < .001$ in all cases), with large effect sizes (η^2p ranging from .16 to .21). In addition, they showed higher levels of perceived greenwashing, confirming their negative impact on message legitimacy.
SO3. Evaluate the potential of Community Notes as mechanisms of epistemic verification.	Community Notes reinforce credibility in truthful messages and correct erroneous perceptions in ambiguous messages.	H3. Hybrid verification systems, such as Community Notes, generate truth cues (cognitive signals that guide the evaluation of message veracity) capable of improving public evaluation of sustainability-related messages, both by identifying greenwashing and reinforcing authentic communication.	Confirmed. The presence of Community Notes increased credibility in authentic messages (M = 6.18 vs. 5.82) and significantly reduced credibility under greenwashing conditions (M = 3.21 vs. 3.94), showing a significant interaction effect, $F(1, 224) = 10.21$, $p = .002$, $\eta^2p = .04$. It also increased perceived greenwashing in inconsistent messages and reduced purchase intention, acting as a mechanism of epistemic amplification and correction.
SO4. Integrate findings to propose an AI + Community model of sustainable credibility.	Credibility emerges from the interaction between algorithmic transparency, community verification, and user sensitivity.	H4. The credibility assigned to an environmental message is significantly influenced by the interaction between factors such as the algorithmic transparency of the communication process, the quality and neutrality of community verification, and the user's prior sensitivity to sustainability, without excluding the potential influence of other contextual and cognitive factors not analyzed in this study.	Confirmed. The factorial design revealed significant interaction effects between message veracity and the presence of Community Notes across all dependent variables analyzed. These results show that credibility emerges from the interaction between content, verification, and interpretative context, confirming the AI + Community model as an explanatory framework for the construction of veracity in digital environments.

Source: Own elaboration based on the integrated results of the three phases of the study (qualitative analysis, survey, and experiment).

7.2. *Answers to the research questions*

The findings allow the research questions to be answered in an integrated and empirically grounded manner. First, evidence across the three studies shows that artificial intelligence applied to persuasive communication exerts an ambivalent effect on the credibility of environmental messages. Algorithmic automation improves communicative effectiveness only when it is perceived as transparent, traceable, and subject to meaningful oversight. This pattern, identified in Studies 1 and 2, is reinforced by the experiment: authentic messages accompanied by verification cues such as Community Notes are perceived as more credible, whereas the absence of such cues intensifies skepticism and perceptions of artificiality (Longoni et al., 2019; Afroogh et al., 2024).

ISSN 2386-7876 – © 2026 Communication & Society, 39(1), 445-463

Secondly, the results confirm that the lack of traceability information and narrative ambiguity erode the epistemic legitimacy of sustainable brands. The perception of greenwashing increases when messages lack verifiable evidence or are perceived as exaggerated, even in previously legitimized brands. In line with Fricker (2007), these findings show that limiting the receiver's cognitive resources constitutes a form of epistemic injustice, and that, in the digital context, narrative traceability stands as the determining factor of credibility, above prior reputation.

Third, the study evidences the value of hybrid models of collaborative verification. Community Notes act as modulatory signals that strengthen trust in truthful messages and activate critical evaluation of doubtful content, confirming their role as tools of communicative governance (Gao et al., 2024; Miceli & Posada, 2022). More than simply substituting for the brand, these mechanisms redistribute epistemic authority to participatory schemes that strengthen users' active roles in interpreting automated content.

Finally, the results show that credibility emerges from a relational architecture in which personalization, verification, and interpretive context interact. AI can optimize communication, but it only builds trust when users have clear signals about its functioning and purpose. Community verification serves as a symbolic moderator, adjusting message perception based on coherence and reception context. This finding reinforces hypothesis H4 and connects with approaches to relational credibility and distributed epistemic justice (Afroogh et al., 2024; Floridi et al., 2018). Taken together, judgments about sustainable messages do not depend solely on their content, but on the conditions under which they are produced, verified, and interpreted.

7.3. *The AI + Community conceptual model: scientific and professional contributions*

This study contributes to the field of digital communication and corporate sustainability by proposing a hybrid interpretive model that articulates artificial intelligence, narrative traceability, and collaborative verification. In a media ecosystem in which algorithmic automation is redefining the codes of persuasion, the AI + Community model offers a framework for sustaining credibility amid greenwashing, information saturation, and growing epistemic distrust.

From a scientific perspective, the model is supported by three convergent dimensions: algorithmic personalization, which enhances communicative effectiveness; symbolic traceability, which is necessary to sustain the legitimacy of green claims; and community verification, understood as a form of distributed epistemic mediation that guides users' critical interpretation. The integrated analysis of the three studies shows that credibility cannot be conceived as a static attribute of either the sender or the content, but rather as a dynamic outcome of the interaction among AI, users, platforms, and interpretive contexts. This approach engages with emerging theories of participatory algorithmic governance (Gao et al., 2024; Raji et al., 2020) and provides empirical evidence relevant to sustainable marketing.

In the professional realm, the findings have direct implications for different actors of the communicative ecosystem:

- For brands, it is evident that technological sophistication alone does not guarantee credibility unless it is accompanied by explainability, verifiable traceability, and human oversight. Signals such as Community Notes, digital passports, or verifiable links can become competitive advantages for those who aim to communicate sustainability with legitimacy.
- For communication professionals: the model redefines the strategic role of "green" messages: it is no longer enough to design effective content, but to build narrative architectures that integrate algorithmic transparency, discursive coherence, and collective verification. This requires new competencies in automation ethics and evidence curation.

ISSN 2386-7876 – © 2026 Communication & Society, 39(1), 445-463

- For digital platforms, the results show that community verification improves the informational quality of the environment, protects users, and reinforces trust in automated content, especially in sensitive sectors like sustainability.
- For regulators: the study reinforces the need to move toward normative frameworks that recognize algorithmic traceability and collaborative verification as essential principles of communicative transparency in the era of automated greenwashing.

In sum, the AI + Community model provides a roadmap for developing more accountable, trustworthy, and sustainable communication systems. Its central contribution lies in articulating the relationships among automated processes and human judgment, persuasive performance and verifiable evidence, and technical infrastructures and symbolic legitimacy. This convergence is essential for advancing toward a digital ecosystem in which sustainable credibility is not merely aspirational, but institutionally and communicatively supported.

8. Limitations and future research

This study provides relevant evidence on the relationships among artificial intelligence, credibility, and community verification, though it has some limitations. First, the experiment used a non-representative sample (initial n = 240; valid n = 228) of regular social media users exposed to sustainability-linked messages, which favors ecological validity but limits generalizability to other contexts.

Second, the use of simulated posts allowed control over the experimental variables, although it reduced the emotional and contextual complexity inherent to real environments. Likewise, the design of Community Notes, even while following neutrality and verifiability criteria, may generate different effects depending on unexamined individual variables, such as epistemic skepticism or ideological predispositions (Longoni et al., 2019; Gunning & Aha, 2019).

Finally, the study is cross-sectional, so it does not capture the evolution of credibility over time nor the cumulative effects of exposure to automated messages or verification mechanisms.

Based on these limitations, several future research directions are proposed: expand the analysis to other sectors, incorporate psychological and cultural variables that help explain individual differences, develop longitudinal designs, and analyze the scalability and governance of systems like Community Notes in relation to emerging regulatory frameworks.

9. Contributions of the study

The present work contributes to the field of digital communication and AI-based marketing by integrating three traditionally separate research streams: algorithmic personalization, credibility in sustainability communication, and hybrid verification models, synthesizing their contributions in theoretical, methodological, and applied terms.

9.1. *Theoretical contributions*

From a theoretical perspective, this study proposes an integrative approach that connects artificial intelligence in persuasive communication (Davenport et al., 2020; Kumar et al., 2021), greenwashing and its impact on trust (Delmas & Burbano, 2011; Santos et al., 2024), and information verification models (Graves, 2016; Lewandowsky et al., 2020).

Additionally, the AI + Community model is introduced as a conceptual framework for analyzing credibility in digital environments, demonstrating that the combination of automation and distributed participation shapes new validation dynamics, particularly in sustainability contexts.

9.2. *Methodological contributions*

From a methodological standpoint, the study contributes a sequential mixed-methods design that integrates theoretical review, reinterpretation of prior studies, and experimental testing through a

ISSN 2386-7876 – © 2026 Communication & Society, 39(1), 445-463

2×2 factorial design. This approach enables the joint analysis of the interaction between message veracity and community-based verification, overcoming the limitations of purely descriptive studies.

The inclusion of an experiment based on simulated social media scenarios, along with manipulation checks, reliability analyses, and a factorial ANOVA, strengthens internal validity and analytical rigor.

9.3. *Practical contributions*

From an applied perspective, the findings offer implications for multiple stakeholders. For brands, credibility depends not only on content but also on its traceability and verifiability, with systems such as Community Notes acting as amplifiers of truthful messages and correctors of greenwashing. For digital platforms, the study provides evidence on the effectiveness of hybrid verification models.

Finally, for consumers and regulators, the results highlight the importance of mechanisms that not only detect misinformation but also facilitate its understanding, reducing interpretive ambiguity and strengthening trust in complex contexts such as sustainability.

10. Acknowledgements

The authors would like to express their gratitude for the support received during the review and preparation of this manuscript, as well as for the conceptual and methodological contributions that have enriched its development.

References

Afroogh, S., Akbari, A., Malone, E., Kargar, M., & Alambeigi, H. (2024). Trust in AI: Progress, challenges, and future directions. *Humanities and Social Sciences Communications, 11*, 1568. https://doi.org/10.1057/s41599-024-04044-8

Bryman, A. (2016). *Social research methods* (5th ed.). Oxford University Press.

Choi, W., Zhu, L., & Lee, H. (2025). Developing an information credibility scale for social media and AI-generated content: Insights from expert and user reviews. *Proceedings of the Association for Information Science and Technology, 62*(1), 1396–1398. https://doi.org/10.1002/pra2.1415

Cohen, S. (1988). Perceived stress in a probability sample of the United States. In S. Spacapan & S. Oskamp (Eds.), *The social psychology of health* (pp. 31–67). Sage.

Couldry, N., & Mejías, U. A. (2019). *The costs of connection: How data is colonizing human life and appropriating it for capitalism*. Stanford University Press.

Creswell, J. W., & Plano Clark, V. L. (2017). *Designing and conducting mixed methods research* (3rd ed.). SAGE.

Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science*, 48, 24–42. https://doi.org/10.1007/s11747-019-00696-0

Delmas, M. A., & Burbano, V. C. (2011). The drivers of greenwashing. *California Management Review*, 54(1), 64–87. https://doi.org/10.1525/cmr.2011.54.1.64

Drolsbach, C., Solovev, K., & Pröllochs, N. (2024). Community Notes increase trust in fact-checking on social media. *PNAS Nexus, 3*(6), 217. https://doi.org/10.1093/pnasnexus/pgae217

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

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, 12 July 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Fernández-Fernández, J. L., & Pinillos, A. (2011). De la RSC a la sostenibilidad corporativa: Una evolución necesaria para la creación de valor (*From CSR to corporate sustainability: A necessary evolution for value creation*). *Harvard Deusto Business Review*, 204, 50–57.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

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

Gao, Y., Zhang, M., & Rui, H. (2024). *Can crowdchecking curb misinformation? Evidence from Community Notes* [Preprint]. SSRN. https://doi.org/10.2139/ssrn.4992470

Graves, L. (2016). *Deciding what's true: The rise of political fact-checking in American journalism*. Columbia University Press.

Gil García, A., & Presol Herrero, Á. (2026). Comunicación persuasiva con IA: personalización, confianza y retos éticos. *index. Comunicación, 16*(1), 37–69. https://doi.org/10.62008/ixc/16/01Comuni

Gunning, D., & Aha, D. W. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI Magazine*, *40*(2), 44–58. https://doi.org/10.1609/aimag.v40i2.2850

Jia, C., & Lee, T. (2024). Journalistic interventions matter: Understanding how Americans perceive fact-checking labels. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-139

Jian, W., & Zhong, Z. (2025). Eco positioning drives sustainable fashion consumption through process-related strategies and brand familiarity. *Scientific Reports, 15*, 17600. https://doi.org/10.1038/s41598-025-02233-2

Koch, T., & Denner, N. (2025). Different shades of green deception: Greenwashing's adverse effects on corporate image and credibility. *Public Relations Review, 51*(1), 102521. https://doi.org/10.1016/j.pubrev.2024.102521

Kumar, V., Ramachandran, D., & Kumar, B. (2021). Influence of new-age technologies on marketing: A research agenda. *Journal of Business Research*, 125, 864–877. https://doi.org/10.1016/j.jbusres.2020.01.007

La Barbera, D., Maddalena, E., Soprano, M., Roitero, K., Demartini, G., Ceolin, D., Spina, D., & Mizzaro, S. (2024). Crowdsourced fact-checking: Does it actually work? *Information Processing & Management, 61*(5), 103792. https://doi.org/10.1016/j.ipm.2024.103792

Lewandowsky, S., Cook, J., Ecker, U. K. H., Albarracín, D., Amazeen, M. A., Kendeou, P., Lombardi, D., Newman, E. J., Pennycook, G., Porter, E., Rand, D. G., Rapp, D. N., Reifler, J., Roozenbeek, J., Schmid, P., Seifert, C. M., Sinatra, G. M., Swire-Thompson, B., van der Linden, S., Vraga, E. K., Wood, T. J., & Zaragoza, M. S. (2020). *The Debunking Handbook 2020*. https://doi.org/10.17910/b7.1182

Lloyd, T., Nguyen, T., Levy, K., & Naaman, M. (2025). *Beyond Community Notes: A framework for understanding and building crowdsourced context systems* [Preprint]. arXiv.

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research, 46*(4), 629–650. https://doi.org/10.1093/jcr/ucz013

Lyon, T. P., & Montgomery, A. W. (2015). The means and end of greenwash. *Organization & Environment, 28*(2), 223–249.

Matassi, M., & Boczkowski, P. J. (2020). Redes sociales en Iberoamérica: Artículo de revisión. *El profesional de la información, 29*(1), e290104. https://doi.org/10.3145/epi.2020.ene.04

Miceli, M., & Posada, J. (2022). The data-production dispositif: Outsourced data work, precarization and knowledge production in Latin American AI pipelines [Preprint]. *arXiv.* https://doi.org/10.48550/arXiv.2205.11963

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2). https://doi.org/10.1177/2053951716679679

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

Pilarski, M., Solovev, K., & Pröllochs, N. (2023). Community Notes vs. Snopes: How the crowd selects fact-checking targets on social media [Preprint]. *arXiv.* https://doi.org/10.48550/arXiv.2305.09519

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33–44). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372873

Santos, C., Coelho, A., & Marques, A. (2024). The greenwashing effects on corporate reputation and brand hate, through environmental performance and green perceived risk. *Asia-Pacific Journal of Business Administration, 16*(3), 655–676. https://doi.org/10.1108/APJBA-05-2022-0216

Sebastião, S., & Dias, D. (2025). AI transparency: A conceptual, normative, and practical frame analysis. *Media and Communication, 13*, Article 9419. https://doi.org/10.17645/mac.9419

Sendra-Duro, E. (2025). *Fact-checking* y desinformación climática en España. Tendencias en la cobertura, tematización y gestión de fuentes en la estrategia profesional de EFE Verifica, Maldito Clima y Newtral. *Doxa Comunicación. Revista Interdisciplinar De Estudios De Comunicación Y Ciencias Sociales*, *41*, 561-587. https://doi.org/10.31921/doxacom.n41a2861

Soundarya, S. (2025). Consumer skepticism towards green marketing: Strategies to build trust. *International Journal of Scientific Research in Engineering and Management, 9*, 1–9. https://doi.org/10.55041/IJSREM42151

Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of human–AI interaction. *Journal of Computer-Mediated Communication, 25*(1), 74–88. https://doi.org/10.1093/jcmc/zmz026

Universitat Oberta de Catalunya. (2023, 21 de diciembre). Ocho de cada diez españoles desconfían de la sostenibilidad empresarial: El peligro del greenwashing. https://n9.cl/pwmrv1

Zarouali, B., Boerman, S. C., Voorveld, H. A. M., & van Noort, G. (2022). The algorithmic persuasion framework in online communication: Conceptualization and a future research agenda. *Internet Research, 32*(4), 1076–1096. https://doi.org/10.1108/INTR-01-2021-0049

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.

Appendices

Appendix 1

AI in Persuasive Communication: Data and Figures. https://doi.org/10.6084/m9.figshare.32256849

Appendix 2

AI, Truth, and Sustainability in Aspirational Brands: Data and Figures
https://doi.org/10.6084/m9.figshare.32256924

Appendix 3

Experiment on Community Notes and Sustainable Credibility: Stimuli and Supplementary Material. https://doi.org/10.6084/m9.figshare.32256930