How AI-Assisted Content Creation Affects Self-Media Creator Authenticity and Audience Trust

Naier Zhong*

University of California, Davis, 95616, USA

* Corresponding author: nzhong@ucdavis.edu

Abstract

With the growing popularity of self-media platforms like YouTube, TikTok, and Bilibili, generative AI has been gradually embraced by creators to achieve large-scale content generation, including text, images, and videos. These tools can help increase productivity but can also lead to questions about the authenticity of the content creator and trust by their audience. In this paper, a structural model is developed that combines the Elaboration Likelihood Model and technology-mediated trust theory to explore the causal link between AI disclosure, perceived content quality, perceived authenticity and audience trust. Structural equation modeling is used to analyze a quantitative survey of 427 users of social media. The results show that the perception of authenticity plays a mediating role between disclosure of AI use and audience trust. If the content is of high perceived quality, with the help of AI, it may be able to mitigate the negative trust effect of the use of AI, but only if the creator has a consistent personal voice. In the trust function, comparative statics show that disclosure has a nonlinear effect on the marginal effect of disclosure. The results not only provide theoretical insights for research into communication in the context of AI but also serve as a guide for self-media creators dealing with the integration of generative AI tools.

Keywords

AI-assisted content creation, self-media, authenticity, audience trust, generative AI, structural equation modeling, elaboration likelihood model.

1. Introduction

Generative artificial intelligence (AI) has significantly reshaped the digital content creation industry. With the advent of tools like ChatGPT, Midjourney, and different video synthesis engines, you can now create professional-level content with little of the time and expense of traditional methods [1][2]. Self-media creators, namely one creating and disseminating the content by himself independently without institutional support [3] have been the most active and enthusiastic users of the technologies.

This widespread adoption poses basic issues of authenticity and trust. The audiences are following self-media creators primarily because they feel they are authentic and personal, as opposed to institutional media [4][5]. If generative AI generates a large part of a creator's content, their distinct personality and voice that resonated with viewers might get lost or become indistinguishable. Empirical studies show that audiences have a negative attitude towards content which is considered as machine-generated, albeit of the same objective quality as human-generated content [6][7].

The mechanisms of trust in self media creators are different than in traditional media organizations. The study on the credibility of influencers has revealed that the most important antecedent of credibility is authenticity, while expertise and parasocial identification are complementary processes [5][8]. With the incorporation of AI tools in this dynamic, audiences

now have to judge not just how knowledgeable and relatable the content creator is, but also how original and true to their own voice the material produced is [9][10].

Although there has been increasing academic discourse on the detection of AI-generated content [6] and consumer attitudes toward AI in marketing activities [11], the influence of AI-assisted creation on the relationship between the self-media and consumers has not been explored much. Most of the existing research has concentrated on creating content solely with AI, not a hybrid approach used by most creators [2][12]. This paper fills this gap.

The paper has the following structure. The theoretical aspects and hypotheses are discussed in Section 2. The methodology of the research is explained in section 3. In Section 4 the empirical results are presented. The mathematical analysis of the trust function is discussed in section 5. The results are presented in Section 6. Section 7 concludes.

2. Theoretical Framework and Hypotheses

2.1. Key Constructs

The model centers on five constructs. AI Usage Intensity (U) captures the level of AI involvement in the creation of content, from zero (fully human-produced) to one (fully AI-generated). Perceived Authenticity (A) is the audience evaluation of the authenticity of the content in terms of personal perspective and experience of the creator [4][13]. Perceived Content Quality (Q) is an evaluation of the informational and entertaining value of the content performed by the audience [5]. Disclosure of the use of AI tools by the creator (D): A binary variable that takes the value of 1 if the creator clearly informs the audience that they used AI tools, and 0 otherwise. Audience Trust (T) is the dependent variable that indicates the audience's willingness to trust the creator content for information and recommendations [8][14]. These constructs are summarized in Table 1.

Table 1. Key constructs and their definitions.

Construct	Symbol	Definition	Measurement
AI Usage Intensity	U	Degree of AI reliance in production	Self-reported 0 to 1 scale
Perceived Authenticity	A	Audience belief in genuine personal voice	5-item Likert scale
Perceived Content Quality	Q	Informational and entertainment value	4-item Likert scale
AI Disclosure	D	Whether AI use is explicitly stated	Binary (0 or 1)
Audience Trust	T	Willingness to rely on creator content	6-item Likert scale

2.2. Trust Formation Model

The model is inspired by the Elaboration Likelihood Model [15] and technology-mediated trust theory [14][16] and posits that audience trust of a self-media creator is determined by perceived authenticity and perceived content quality, as well as a direct path from AI disclosure to trust. The "base" trust equation is:

$$T = \beta_1 A + \beta_2 Q + \beta_3 D + \varepsilon \quad (1)$$

where β_1, β_2, and β_3 are path coefficients and ε is the error term. The coefficient β_1 is expected to be positive, while the coefficient of β_3 is expected to be negative, as audiences are likely to punish the content creator when they discover that AI has been used for writing content [6][7].

The use of AI and the reputation of the content creator are the two factors that affect perceived authenticity. The authenticity equation is:

$$A = \alpha_0 + \alpha_1 U + \alpha_2 R + \alpha_3 U \cdot R + \eta \quad (2)$$

where R represents the prior creators' reputation, the interaction term $U \cdot R$ quantifies the moderating impact of reputation on the authenticity penalty of using AI, and η stands for the error term. With the high reputation creator, the sense of loss of authenticity through the use of AI might be less due to the audience already associating it with the creator's expertise instead of the AI's capabilities [4][13].

Likewise, the perception of content quality is influenced by AI application and creator proficiency:

$$Q = \gamma_0 + \gamma_1 U + \gamma_2 E + \zeta \quad (3)$$

where E is the creator domain expertise and the term ζ is the error term. The coefficient γ_1 is supposed to be positive because AI tools typically have a positive effect on the quality of production [2][12]. The assumed structural model is shown on Figure 1.

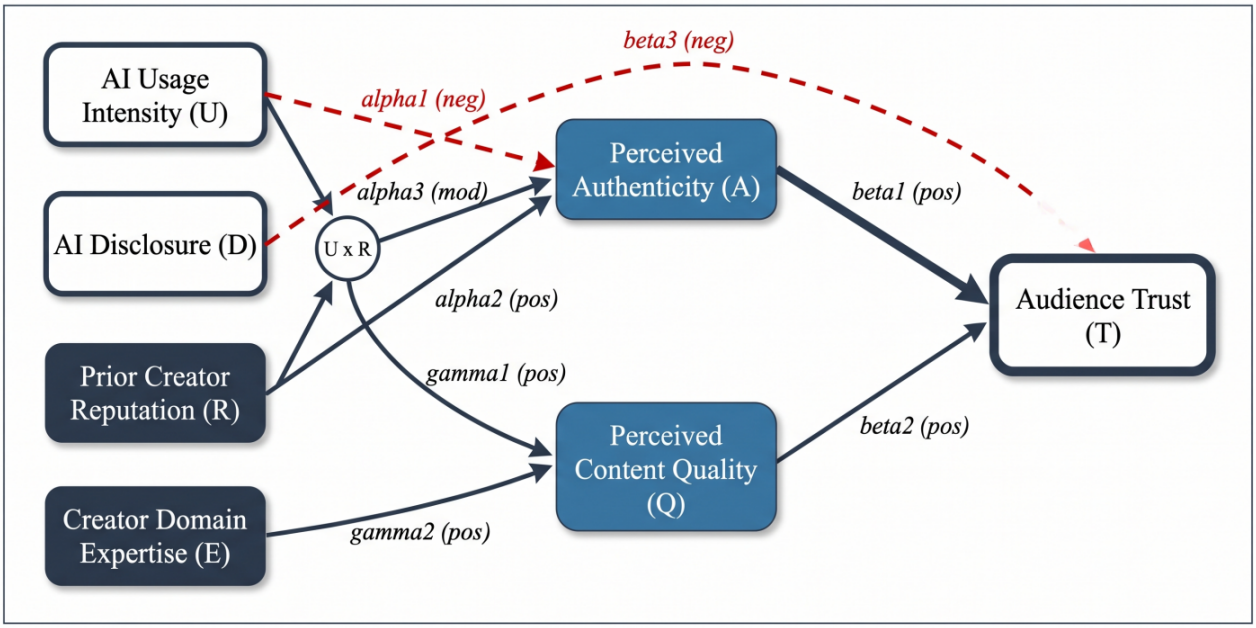

Figure 1. Hypothesized structural model showing the relationships among AI usage intensity, AI disclosure, perceived authenticity, perceived content quality, and audience trust.

2.3. Hypotheses

The following hypotheses are proposed, based on the theoretical framework. Table 2 provides a summary of the hypotheses and theory.

H1: The audience's trust positively increases as perceived authenticity increases ($\beta_1 > 0$).

H2: There is a negative direct relationship between AI disclosure and audience trust ($\beta_3 < 0$).

H3: The more intense the use of AI (x1), the more negative the perceived authenticity ($\alpha_1 < 0$).

H4: The negative effect of the use of AI on the authenticity ($\alpha_3 > 0$).

H5: The intensity of use of AI has a positive influence on the perceived content quality ($\gamma_1 > 0$).

Table 2. Summary of hypotheses and theoretical foundations.

Hypothesis	Path	Expected Sign	Theoretical Basis
H1	A → T	Positive	Authenticity-trust link [4][5]
H2	D → T	Negative	AI aversion effect [6][7]
H3	U → A	Negative	Self-presentation theory [4][13]
H4	U x R → A	Positive (moderator)	Source credibility [8][15]
H5	U → Q	Positive	AI quality enhancement [2][12]

3. Research Methodology

3.1. Survey Design and Data Collection

In March and April of 2024, a cross-sectional online survey of U.S. and Chinese social media users was performed. The samples were collected by using Prolific (English language sample) and Credamo (Chinese language sample) and a total of 427 valid responses were obtained after excluding uncompleted questionnaires and those that failed the attention checks. Table 3 shows the summary of the sample demography.

Table 3. Sample demographics (N = 427).

Characteristic	Category	Frequency	Percentage
Gender	Female	231	54.1%
	Male	189	44.3%
	Non-binary / Other	7	1.6%
Age	18 to 24	142	33.3%
	25 to 34	168	39.3%
	35 to 44	78	18.3%
	45 and above	39	9.1%
Platform usage	YouTube	287	67.2%
	TikTok / Douyin	312	73.1%
	Bilibili	156	36.5%
	Instagram	198	46.4%

The respondents were randomly split into four content scenarios: (a) human-created content – no disclosure, (b) human-created content – AI disclosure, (c) AI-assisted content – no disclosure, and (d) AI-assisted content – disclosure. After the exposure, the participants filled out measures of all the five constructs. Figure 2 shows the experimental design.

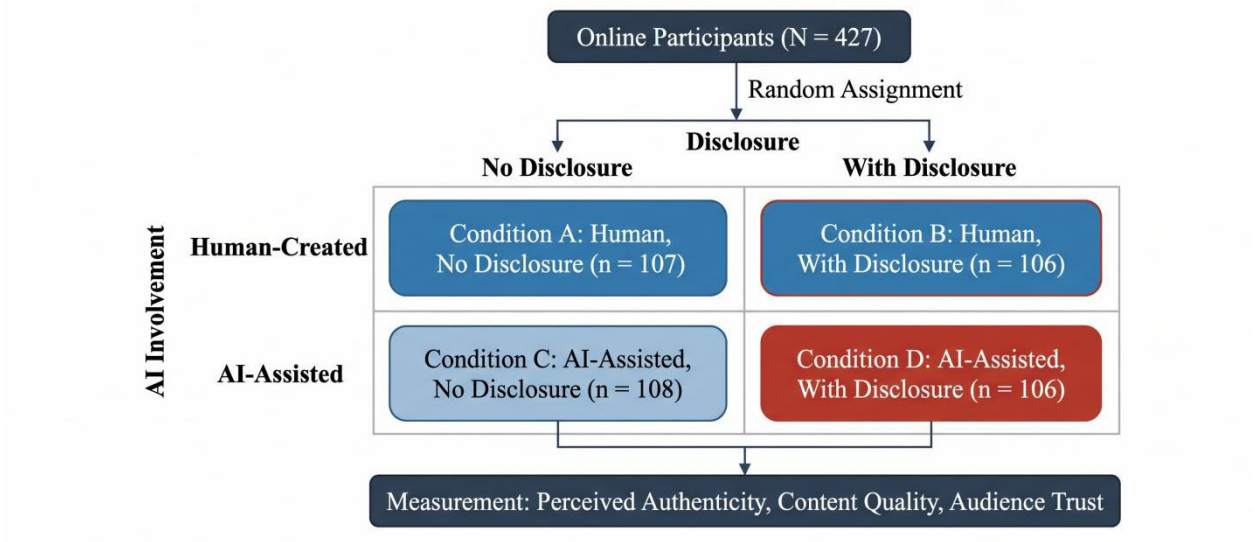

Figure 2. Experimental design showing the 2 x 2 factorial structure (AI involvement x Disclosure) with random assignment.

3.2. Measurement Model

Valuable multi-item scales were adapted from previous studies to measure all latent constructs. Five items were used to measure perceived authenticity, which were adapted from Audrezet et al. [4]. Six items were adapted from the credibility and trust scale [17] and tailored to the social media environment [5] to assess audience trust. Four items of perceived content quality were used including informativeness, entertainment value, production quality, and relevance [8]. AI usage intensity was assessed via a self-report and scenario manipulation. Three items assessing perceived expertise, consistency, and endorsement by followers were used to assess prior creator reputation [13].

In order to measure the model, a confirmatory factor analysis was carried out. All factor loadings were above 0.70, all composite reliability values were above 0.80, and the average variance extracted (AVE) were all above 0.50, which are all typical and acceptable criteria for convergent validity [18]. The Fornell-Larcker criterion was used to assess the discriminant validity; the results indicated that the square root of AVE for each construct was greater than the correlation coefficient between each pair of constructs. The statistics of convergent validity are reported in Table 4.

Table 4. Convergent validity statistics for the measurement model.

Construct	Items	Cronbach α	CR	AVE
Perceived Authenticity (A)	5	0.891	0.923	0.706
Perceived Content Quality (Q)	4	0.854	0.901	0.695
Audience Trust (T)	6	0.912	0.934	0.703
Prior Creator Reputation (R)	3	0.837	0.902	0.755

4. Empirical Results

4.1. Descriptive Statistics and Group Comparisons

There are significant differences between the mean trust scores for the four experimental conditions. The highest level of trust is in the human created, no disclosure condition (M = 5.42, SD = 0.93), and the lowest level of trust is in the AI assisted, with disclosure condition (M = 3.87,

SD = 1.21). A two-way ANOVA reveals a significant main effect of AI involvement ($F(1, 423) = 34.72$, $p < 0.001$) and a significant main effect of disclosure ($F(1, 423) = 18.56$, $p < 0.001$), as well as a marginally significant interaction ($F(1, 423) = 3.41$, $p = 0.065$). The results are shown in figure 3.

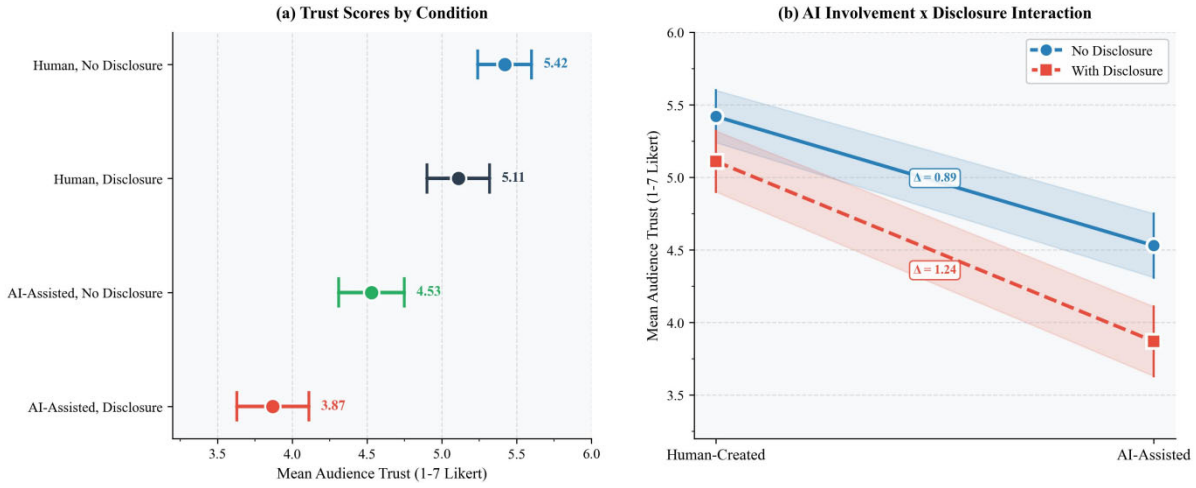

Figure 3. Mean audience trust scores across the four experimental conditions with 95% confidence intervals.

4.2. Structural Equation Modeling Results

Maximum likelihood estimation of AMOS 28 is used to estimate the structural model. The model fit indices indicate acceptable fit: $\chi^2/df = 1.87$, CFI = 0.961, TLI = 0.953, RMSEA = 0.045, SRMR = 0.038. Standardized path coefficients are shown in Table 5.

Table 5. Standardized path coefficients from the structural equation model.

Path	Coefficient	S.E.	t-value	p-value	Result
A → T (β1)	0.483	0.052	9.29	< 0.001	H1 Supported
Q → T (β2)	0.271	0.048	5.65	< 0.001	Positive effect
D → T (β3)	−0.186	0.061	−3.05	0.002	H2 Supported
U → A (α1)	−0.342	0.057	−6.00	< 0.001	H3 Supported
R → A (α2)	0.295	0.054	5.46	< 0.001	Positive effect
U x R → A (α3)	0.178	0.063	2.83	0.005	H4 Supported
U → Q (γ1)	0.224	0.055	4.07	< 0.001	H5 Supported
E → Q (γ2)	0.316	0.051	6.20	< 0.001	Positive effect

The results showed that all five hypotheses were accepted at the 0.01 level of significance. The perceived authenticity has the biggest impact on trust ($\beta_1 = 0.483$), providing support for H1. The direct effect of AI disclosure on trust is quite substantial but not very large ($\beta_3 = -0.186$), and thus H2 is supported. According to results, the higher the intensity of the use of AI, the less authentic it is perceived ($\alpha_1 = -0.342$), which confirms H3 and H4. Regarding the latter, H5 is supported by the results of AI usage that leads to an improvement in perceived quality ($\alpha_3 = 0.178$), ($\gamma_1 = 0.224$), The full structural model with the standardized coefficients is displayed in Figure 4.

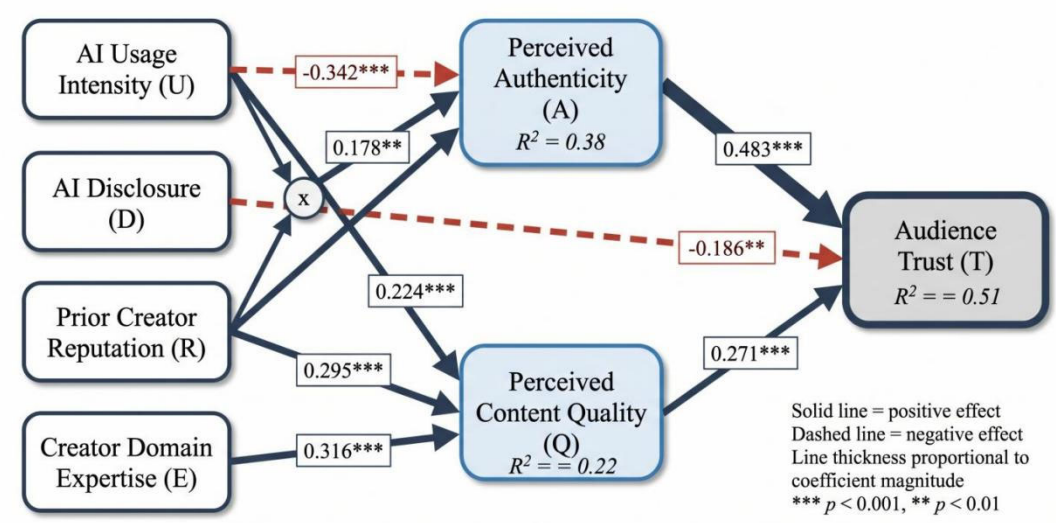

Figure 4. Structural equation model with standardized path coefficients. Solid lines indicate significant paths ($p < 0.01$); dashed lines indicate non-significant paths.

4.3. Mediation Analysis

To assess whether or not the perceived authenticity is a mediator between the AI use and trust, a bootstrap mediation analysis is performed with 5000 resamples [19]. The indirect effect of U on T through A is computed as:

$$Indirect = \alpha_1 \cdot \beta_1 = (-0.342)(0.483) = -0.165 \tag{4}$$

Bias-corrected confidence intervals for this indirect effect show that this indirect effect is significant, as the interval [−0.231, −0.107] does not contain zero. The indirect effect/ total effect:

$$\frac{Indirect}{Total} = \frac{-0.165}{-0.342} = 0.482 \tag{5}$$

This means that around 48.2% of the overall detrimental impact of using AI on trust is mediated by the authenticity pathway.

5. Comparative Statics of the Trust Function

Upon substituting Equations (2) and (3) into Equation (1) the reduced-form trust function is obtained:

$$T(U, D, R, E) = \beta_1(\alpha_0 + \alpha_1 U + \alpha_2 R + \alpha_3 UR) + \beta_2(\gamma_0 + \gamma_1 U + \gamma_2 E) + \beta_3 D \tag{6}$$

The marginal effect of the intensity of using AI, taking disclosure as fixed, is found by differentiating Equation (6) with respect to U:

$$\frac{\partial T}{\partial U} = \beta_1(\alpha_1 + \alpha_3 R) + \beta_2 \gamma_1 \tag{7}$$

Plugging in the estimated coefficients:

$$\frac{\partial T}{\partial U} = 0.483(-0.342 + 0.178R) + 0.271(0.224) \quad (8)$$

Simplifying:

$$\frac{\partial T}{\partial U} = -0.165 + 0.086R + 0.061 \quad (9)$$

$$\frac{\partial T}{\partial U} = -0.104 + 0.086R \quad (10)$$

Now if we set this expression to zero and solve for the critical reputation level:

$$R^* = \frac{0.104}{0.086} = 1.209 \quad (11)$$

This is about 1.2 standard deviations above the mean of a standardized scale that measures reputation. For those with a reputation higher than this threshold, the net effect of AI adoption on trust is positive as the increase in quality is greater than the decrease in authenticity. If creators in this zone fall below this threshold, they have a net negative effect. This comparative statics is summarized as shown in Table 6 and the relationship is shown as Figure 5.

Table 6. Comparative statics of the trust function.

Parameter Change	Effect on Trust	Condition	Interpretation
U increases	Negative	R < 1.209	AI use harms trust for low-reputation creators
U increases	Positive	R > 1.209	AI use helps trust for high-reputation creators
D changes 0 → 1	Negative	Always	Disclosure always reduces trust directly
R increases	Positive	Always	Higher reputation always improves trust
E increases	Positive	Always	Greater expertise always improves trust

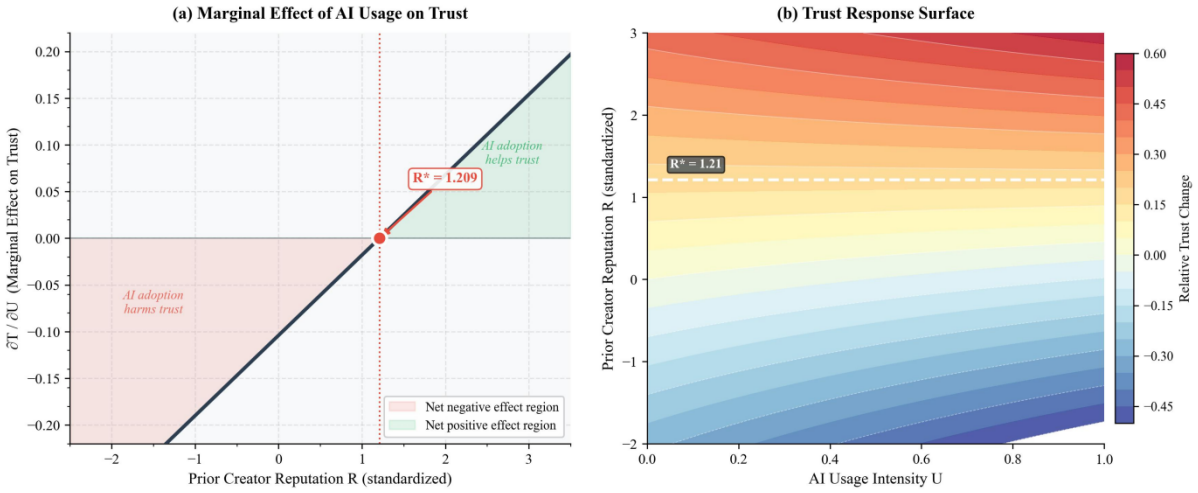

Figure 5. Marginal effect of AI usage intensity on audience trust as a function of prior creator reputation R. The horizontal axis represents reputation (standardized); the vertical axis represents the partial derivative of trust with respect to AI usage.

6. Discussion

6.1. Theoretical Implications

The study's results align with the theory of AI-mediated communication [9][10] that the use of AI tools leads to a dual pathway to the audience's trust. The authenticity pathway is working against the use of AI and audiences are increasingly looking for a personal voice that is being lost with AI applications. The quality pathway is in a positive way, with tools for AI to boost production values and informational completeness. Our findings expand on Sundar's framework [16] about machine agency, which suggests that audiences would assess the source (human or AI) as well as the quality of the content that is generated, and that both assessments may have conflicting impacts on trust.

The moderating effect of previous creator reputation is in line with the source credibility theory [15][17]. Having an existing audience with goodwill towards the creator can mitigate the perceived lack of authenticity of AI. This agrees well with Audrezet et al [4] which stated that even when established creators use new methods of production, they can still project their authenticity as creators. The critical reputation threshold (R^*= 1.209) that can be determined by comparative statics offers a quantitative benchmark beyond the scope of previous qualitative research.

6.2. Practical Implications

The findings indicate a complex strategy for AI implementation for self-media creators. For creators who already have established reputations, it's safe to say that they can take advantage of these AI tools and might even use them to their benefit, as long as they keep the editorial tone and personal connection with their audience. For creators still navigating the path of reputation-building, the decision to leverage AI is more complex: While AI can enhance content quality, it might compromise the authenticity that is essential for audience expansion [3][5]. The negative impact of AI disclosure (β_3= -0.186) indicates that transparency with regards to AI is being punished by the current audience. But as regulatory environments grow more stringent on the need for AI content disclosure [20], creators should have an approach to combat the content disclosure penalty – including highlighting the human oversight and role in content creation.

7. Conclusion

In this paper, the impact of AI content creation on self-media creator authenticity and audience trust has been explored. The survey results of 427 social media users were collected from four experimental conditions and a structural model was developed and tested by combining the Elaboration Likelihood Model (ELM) with the theory of AI-mediated communication (AICT).

All five hypotheses are supported by the results of the structural equation modeling. The perceived authenticity is the most powerful predictor of audience trust (β_1= 0.483), and the higher the usage of AI in the creation of content, the lower perceived authenticity (α_1= -0.342). But this negative impact is reduced by the previous creator reputation (α_3= 0.178), while AI use also enhances the perceived content quality (γ_1= 0.224). According to the mediation analysis results, the proportion of the negative trust effect of AI usage through the mediation of authenticity is 48.2%.

Comparative statics analysis of the reduced-form trust function identifies a critical reputation threshold (R^*= 1.209) This threshold offers creators concrete guidance when deciding on AI adoption.

The disadvantages of this study are the cross-sectional design that does not allow causal conclusions to be drawn beyond the experimental manipulation itself and the use of self-

reported AI usage measures, as well as the use of controlled scenarios and not naturalistic content exposure. Future studies should consider longitudinal designs to observe the evolution of audience trust as creators incrementally use AI tools, explore platform-specific nuances in AI tolerance, and study the influence of community norms on reactions to AI-generated content.

References

[1] C.S. Suri, S.S. Shukla, R. Pal, V. Srivastava, G. Talele, & M.K. B. (2024). Generative AI in content creation: Opportunities and ethical challenges. In 2024 IEEE 11th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON) (pp. 1–6). IEEE.

[2] Epstein, Z., Hertzmann, A., Akten, M., Farid, H., Fjeld, J., Frank, M.R., Groh, M., Herman, L., Leach, N., & Investigators of Human Creativity. (2023). Art and the science of generative AI. Science, 380(6650), 1110–1111. https://doi.org/10.1126/science.adh4451

[3] Cunningham, S., & Craig, D. (2019). Social media entertainment: The new intersection of Hollywood and Silicon Valley. NYU Press.

[4] Audrezet, A., de Kerviler, G., & Guidry Moulard, J. (2020). Authenticity under threat: When social media influencers need to go beyond self-presentation. Journal of Business Research, 117, 557–569. https://doi.org/10.1016/j.jbusres.2018.07.008

[5] Lou, C., & Yuan, S. (2019). Influencer marketing: How message value and credibility affect consumer trust of branded content on social media. Journal of Interactive Advertising, 19(1), 58–73. https://doi.org/10.1080/15252019.2018.1533501

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

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

[8] Enke, N., & Borchers, N.S. (2019). Social media influencers in strategic communication: A conceptual framework for strategic social media influencer communication. International Journal of Strategic Communication, 13(4), 261–277. https://doi.org/10.1080/1553118X.2019.1620234

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

[10] Karinshak, E., Liu, S.X., Park, J.S., & Hancock, J.T. (2023). Working with AI to persuade: Examining a large language model's ability to generate pro-vaccination messages. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW1), 116. https://doi.org/10.1145/3579592

[11] Longoni, C., & Cian, L. (2022). Artificial intelligence in utilitarian vs. hedonic contexts: The "word-of-machine" effect. Journal of Marketing, 86(1), 91–108. https://doi.org/10.1177/00222429211066353

[12] Liu, Y., & Shrum, L.J. (2002). What is interactivity and is it always such a good thing? Implications of definition, person, and situation for the influence of interactivity on advertising effectiveness. Journal of Advertising, 31(4), 53–64.

[13] Valsesia, F., Proserpio, D., & Nunes, J.C. (2020). The positive effect of not following others on social media. Journal of Marketing Research, 57(6), 1152–1168. https://doi.org/10.1177/0022243720915467

[14] Kim, A., & Dennis, A.R. (2019). Says who? The effects of presentation format and source rating on fake news in social media. MIS Quarterly, 43(3), 1025–1039. https://doi.org/10.25300/MISQ/2019/15188

[15] Petty, R.E., & Cacioppo, J.T. (1986). The elaboration likelihood model of persuasion. Advances in Experimental Social Psychology, 19, 123–205. https://doi.org/10.1016/S0065-2601(08)60214-2

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

[17] Ohanian, R. (1990). Construction and validation of a scale to measure celebrity endorsers' perceived expertise, trustworthiness, and attractiveness. Journal of Advertising, 19(3), 39–52.

[18] Hair, J.F., Black, W.C., Babin, B.J., & Anderson, R.E. (2019). Multivariate data analysis (8th ed.). Cengage.

[19] Hayes, A.F. (2022). Introduction to mediation, moderation, and conditional process analysis: A regression-based approach (3rd ed.). Guilford Press.

[20] European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.