Uysal, Busye; Estrella, Ronny

Conference Paper

Navigating Illusions: Unraveling Confirmation Bias using Cognitive Dissonance in Virtual Influencers on Social Media Platforms

24th Biennial Conference of the International Telecommunications Society (ITS): "New bottles for new wine: digital transformation demands new policies and strategies", Seoul, Korea, 23-26 June, 2024

Provided in Cooperation with:
International Telecommunications Society (ITS)

Suggested Citation: Uysal, Busye; Estrella, Ronny (2024) : Navigating Illusions: Unraveling Confirmation Bias using Cognitive Dissonance in Virtual Influencers on Social Media Platforms, 24th Biennial Conference of the International Telecommunications Society (ITS): "New bottles for new wine: digital transformation demands new policies and strategies", Seoul, Korea, 23-26 June, 2024, International Telecommunications Society (ITS), Calgary

This Version is available at:
https://hdl.handle.net/10419/302466

Navigating Illusions: Unraveling Confirmation Bias using Cognitive Dissonance in Virtual Influencers on Social Media Platforms

Busye Uysal[a] , Ronny Estrella[a]

[a]SolBridge International School of Business, 128 Uam-ro, Samseong-dong, Dong-gu, Daejeon, Korea

While policymakers are beginning to address AI-generated content on social media, there remains a notable gap in regulatory approaches towards Virtual Influencers. The capability of Virtual Influencers to autonomously upload content presents significant challenges, especially in distinguishing between human and AI-generated content, which in turn affects user trust. To tackle this issue, this study proposes the implementation of disclosure flags specifically for content created by Virtual Influencers. This research involved a questionnaire administered to 189 Instagram users to explore how disclosure flags influence their perceptions and acceptance of Virtual Influencer's content. The findings reveal that although disclosure flags increase awareness, they do little to foster critical engagement with the content. The study emphasizes the importance of professional oversight and user-driven content moderation through disclosure flags to maintain the integrity of digital content. These insights are crucial for policymakers and platform designers working towards a transparent digital environment. The evident lack of transparency around Virtual Influencers highlights the urgent need for clearer regulatory frameworks. Therefore, this research advocates for comprehensive strategies that integrate these flags with broader educational and regulatory measures to enhance digital literacy and critical engagement among users.

Keywords: Virtual Influencers, Social Robots, Social Media, Cognitive Dissonance, Affective Behavior, Disclosing Flags

1. Introduction

Social media influencers (SMIs) have revolutionized marketing in the digital era, playing a crucial role in shaping consumer decisions and behavior, especially on social media platforms (Antunes, 2022). SMIs are often defined as specialists in personal branding, who cultivate a unique public image visible through their online presence (Djafarova & Trofimenko, 2019; Khamis et al., 2017). These SMIs, recognized as opinion leaders, wield a significant influence on public opinion and consumer behavior (Casaló et al., 2020; Ha & Yang, 2023). Through their authentic and relatable content, SMIs have become trusted sources of information and trendsetters in various industries (Djafarova & Trofimenko, 2019). Their ability to engage with audiences on a personal level and showcase products or services in a genuine manner has disrupted traditional advertising methods (Phua et al., 2017). As a result, businesses are increasingly turning to influencer marketing as a means to reach and connect with their target demographics in a more impactful and authentic way (Freberg et al., 2011). Leveraging SMIs can lead to significant earnings, with some firms

Email addresses: busyeuysal@solbridge.ac.kr (B.Uysal), restrella@solbridge.ac.kr (R.Estrella).

earning $18 in media value for every dollar allocated, while smaller businesses receive an average of $5.78 in media value per dollar allotted (Geyser, 2020). By 2029, the market size of influencer advertising is expected to be valued at around $56.28 billion (Statista, 2024). This projection highlights the sustained expansion and increasing dependency on influencer marketing as a crucial marketing strategy, emphasizing its significant and ongoing impact on the advertising industry.

As the demand for SMIs grows and the market continues to evolve with companies seeking effective digital marketing strategies, a significant transformation has taken place in the field of influencer marketing. This development has led to the emergence of innovative entities referred to as Virtual Influencers (VIs).VIs are computer-generated images (CGI) or animated digital characters (Bringe, 2022), that are designed to resemble humans and mimic various human traits, consequently generating a significant following on various social media platforms (Arsenyan & Mirowska, 2021; Choudhry et al., 2022; Conti et al., 2022). VIs have garnered substantial followings on the Social Networking Site (SNS) Instagram (de Brito Silva et al., 2022). Marketing companies are choosing VIs over human influencers mainly because of their capacity to offer complete control, cost-effectiveness, and adaptability (Muttamimah & Irwansyah, 2023). Unlike SMIs, whose actions and behaviors may be unpredictable, VIs can be programmed to strictly adhere to brand guidelines, mitigating the risk of reputational damage or brand misalignment (Xin et al., 2024). The importance of VIs in marketing is solidified based on the market's rapid growth. The VI market is exponentially expanding, with a yearly growth rate of 37%. In just five years, it jumped from $2.2 billion to $10.8 billion (Premia, 2022). This surge in market size highlights the increasing adoption and effectiveness of VIs as a viable marketing tool in the digital age.

VIs have emerged as opinion leaders that can significantly impact engagement on social media platforms, as they generate more engagement in the form of likes, comments, and word of mouth than institutional influencers (Almeida et al., 2018). Lil Miquela, the pioneering VI developed by the transmedia studio Brud, emerged in 2016 (Parsani, 2023). By 2018, her Instagram following exceeded 1 million users, a figure that has since grown to 2.6 million as of April 2024 (Drenten & Brooks, 2020). She has been named as one of Time's most influential figures on the internet, alongside notable personalities such as Rihanna, Trump, and Kanye West (Time, 2018). Over time, her content has evolved from simple photo uploads to sponsored brand collaboration. Notably, Lil Miquela has established partnerships with prestigious fashion houses such as Chanel, Burberry, and Fendi (Sands et al., 2022). These partnerships underscore the expanding reach of VIs into industries beyond luxury fashion, demonstrating their growing influence and marketability in diverse sectors. One such collaboration is Lil Miquela's partnership with BMW for the promotion of its innovative all-electric iX2 vehicle (Junkie, 2023). This initiative exemplifies the widespread impact and relevance of VIs beyond the confines of traditional SNS, as they increasingly engage in high-profile collaborations with prominent brands, garnering significant attention in mainstream media. The heightened consumer engagement with VIs can be attributed to several factors, including the immersive experience facilitated by CGI and the distinctive aesthetic qualities they embody (Lou et al., 2023). These unique characteristics contribute to the allure and appeal of VIs, fostering deeper connections and interactions with their audience.

The proliferation of VIs has sparked controversy among researchers, with some skeptical of their ethics (Conti et al., 2022; Mertens & Goetghebuer, 2024). Currently, most firms that oversee VIs act primarily as intermediates, developing market strategies, managing the interactions between sponsors and audiences, and supervising communication implementation (Chow, 2023). Moreover, the actions of VIs are firmly linked to their developers or creators, portraying them as tools through which humans exert influence (D. Kim & Wang, 2023). The concept of autonomy is a fundamental element of artificial intelligence (AI), distinguishing advanced technologies from previous generations (Scherer, 2015). As AI advances and

Email addresses: busyeuysal@solbridge.ac.kr (B.Uysal), restrella@solbridge.ac.kr (R.Estrella).

systems demonstrate greater autonomy, the potential of autonomous VIs presents a viable path forward (Mertens & Goetghebuer, 2024). VIs, being AI-driven entities (Sands et al., 2022), may develop independent capabilities similar to those found in advanced AI systems. Thus, there is a discernible inference of the potential for VIs to autonomously upload content on social media platforms independently without needing direct parent company decision-making, human intervention, or supervision (Mertens & Goetghebuer, 2024).

The potential of the autonomy of VIs raises several concerns for users, necessitating governmental intervention to regulate VIs on social media (Mertens & Goetghebuer, 2024). Currently, certain platforms are taking a step forward in the direction of inclusivity by implementing measures to address the growing presence of AI-generated content. Meta's approach involves labeling AI-generated content to inform users and mitigate deception risks (Meta, 2024), while TikTok is empowering creators with tools to label their AI-generated content and testing automated labeling systems (TikTok, 2023). These initiatives reflect a proactive response to the evolving landscape of digital content creation, ensuring users are better informed about the origins and nature of the content they consume. However, these steps primarily rely on user discretion, lacking government mandates.

The prospect of VIs gaining autonomy to upload content poses a challenge for users to discern whether it originates from humans or VIs, thereby blurring the lines between entities and undermining transparency and authenticity on social media platforms, consequently eroding user trust (Mertens & Goetghebuer, 2024). Without governmental requirements, there is no guarantee that VI's profiles or posts will be distinguishable for users, raising concerns about the spread of misinformation and manipulation. The absence of disclosure regarding the AI-driven nature poses challenges for users in discerning potential misinformation, fake news, false endorsements, or manipulated information, thereby increasing the risk of significant societal harm. This includes the propagation of conspiracy theories, divisive narratives, and unethical marketing practices (Mertens & Goetghebuer, 2024). Without the differentiating point of SMIs and VIs presented by social media companies, confirmation bias can occur in the users.

Confirmation bias (Nickerson, 1998) occurs` when users encounter information that reinforces their pre-existing beliefs and attitudes (Bessi, 2016; Geschke et al., 2019; Modgil et al., 2024; Zhao et al., 2020). In the case of VIs, users may interpret the content generated by VIs in a way that aligns with their preexisting beliefs or preferences (Mertens & Goetghebuer, 2024). The issue with this is that some people end up with even more extreme positions even after they actively seek out dissimilar or disagreeable information (Ditto & Lopez, 1992; Taber & Lodge, 2006). Considering the increasing indistinctiveness between human-operated and autonomous VIs, the absence of governmental oversight poses a risk to users, underscoring the importance of clear differentiation for user welfare.

However, from an academic perspective, previous research has intensely focused on the differences between SMIs and VIs and the purchase behavior of sponsored content, with additional factors that positively or negatively impact the source of the messages of VIs (D. Kim & Wang, 2023; Mertens & Goetghebuer, 2024; Sands et al., 2022).

To the best of our knowledge, there has been no prior research into understanding the importance of disclosure or knowledge validation in AI-generated content. The existing research does not encompass factors or mechanisms that allow users to validate the broader spectrum of artificial intelligence-generated content. By broadening our understanding to include knowledge validation, we can better grasp the complexities and implications of disclosure practices in this rapidly advancing field.

To address this problem, this research suggests using a variety of presenting strategies for VIs on social media platforms, encompassing the utilization of 'disclosing' flags. Disclosing flags on SNS is a mechanism for users to report offensive content, acting as an instrument for content monitoring (Crawford & Gillespie, 2016). Flagging the content serves as a solution for organizing large collections of

Email addresses: busyeuysal@solbridge.ac.kr (B.Uysal), restrella@solbridge.ac.kr (R.Estrella).

user-generated information as well as a rhetorical justification for platform administrators when they decide to remove content. Flags are becoming more and more common as a governance and content moderation tool (Lanius et al., 2021). Our study differs from existing research as we present mechanisms and theoretical advancements from the knowledge validation perspective.

This study aims to bridge the aforementioned gaps in the literature of moderating VIs on social media by proposing the introduction of disclosure flags. Grounded in the theory of cognitive dissonance, this approach seeks to define how users currently respond to SNS-imposed disclosure flags for VIs, thus linking these flags as a potential solution to confirmation bias on social media. This new element aims to investigate the effects of transparency and disclosure on users' assessment and acceptance of VIs. Consequently, the findings of this research are expected to contribute to scholarly discussions and provide insights for companies and governments to effectively engage with VIs, ultimately improving outcomes in their respective areas.

The rest of the paper is structured as follows. In sections 2 and 3, the theoretical background regarding confirmation bias, cognitive dissonance theory, and VIs on social media platforms will be described and related hypotheses will have introduced. In section 4, the methodological data collection and analyses will be elaborated. In the last section, the contributions and limitations that can be derived from the quantitative results will be explained.

2. Literature Review and Theoretical Background

This section discusses themes heavily researched in the VI sphere. Subsequently, foundational underpinnings of the confirmation bias theory are introduced, contextualizing its relevance within the field of social media, and explores its applicability to VIs. Furthermore, cognitive dissonance theory is discussed and provides a background of the theory, subsequently establishing its relevance to VIs, particularly how it can arise on SNS.

2.1. Review of prior literature on VIs

Researchers in existing studies have thoroughly investigated the marketing perspective of VIs. The impact of VIs on consumer behavior and marketing effectiveness has emerged as a prominent theme in recent years, garnering growing research attention. One research stream focuses on the effectiveness of VIs in the context of consumer attitudes. Previous studies have extensively investigated engagement (de Brito Silva et al., 2022; Yu et al., 2024) and consumer perception (De Cicco et al., 2024; Jang & Yoh, 2020) in this regard. Visibility (Moustakas et al., 2020), authenticity of appearance (Koles et al., 2024), and brand fit of VIs (H. Kim & Park, 2023), as well as engagement, creativeness, and brand narrative in advertising content design were favorably associated with customer brand engagement on social media platforms (Zhong, 2022). Additionally, maintaining an equilibrium between authenticity and product engagement is crucial for preserving the sense of anthropomorphism and authenticity, thus influencing advertising perceptions (Um, 2023) and implicit actions. Realism and product interaction can enhance impressions of anthropomorphism and authenticity but overbearing integration of reality, like the instance of VIs consuming a real-world branded product alongside a real human in a single social media post undermines these effects (Ham et al., 2023).

The other stream of research focuses on the attractiveness of VIs in the context of business expectations. For marketing firms interested in VIs, they are cost-effective compared to SMIs (Franke et al., 2023). Brands perceive VIs as more controllable in communication strategies, leading to more predictable outcomes in marketing campaigns.

While previous studies' overarching theme revolves around comparing VIs with traditional human SMIs with the primary focus on consumers, it is necessary to redirect the research focus away from the consumer perception and advertising aspect of VIs. For ethical considerations, previous research emphasized that the inability to separate VIs and SMIs raises questions about the ethical construction of identity (Robinson, 2020). Concerns extend to issues of accountability, particularly in scenarios where these

Email addresses: busyeuysal@solbridge.ac.kr (B.Uysal), restrella@solbridge.ac.kr (R.Estrella).

entities might inadvertently endorse harmful content or products (Mertens & Goetghebuer, 2024). A notable deficiency within the existing literature lies in its failure to provide a viable mechanism on SNS for addressing these concerns. It is clear that VIs are valuable for marketing; thus, it is important to consider that ethical considerations, particularly regarding regulations for SNS enforced by governments, and the exploration of potential risks and critical awareness, remain relatively underexplored areas of inquiry.

To address the critical research gaps, this study will draw upon the theory of confirmation bias. This theory will serve as the basis for designing our conceptual and operational framework.

2.2 Theory of Confirmation Bias

Confirmation bias is defined as the tendency to favor information that confirms pre-existing beliefs or hypotheses while disregarding contradictory evidence. Several research studies have identified a link between social media participation and confirmation bias. These studies explain how social media platforms influence the formation and continuation of people's confirmation biases in digital contexts (Ghani & Rahmat, 2023). Confirmation bias is prevalent among partisans, who prefer to seek information that reinforces their political ideas on social media (Rahkman Ardi, 2021). Studies have demonstrated that social media platforms can exacerbate confirmation bias by creating echo chambers and filter bubbles, where users are exposed to content that reinforces their existing viewpoints while shielding them from opposing perspectives or contradictory information. Moreover, the viral spread of misinformation and fake news on social media platforms can exploit confirmation bias, as users are more likely to accept and share content that aligns with their beliefs, regardless of its accuracy (Pennycook & Rand, 2019). Given the ambiguity surrounding the identity of the actual entity or individual behind the VI persona, users' preferences for reinforcing content can significantly influence their perceptions and interactions with VIs. This, in turn, may shape users' attitudes, behaviors, and brand preferences based on content that aligns with their existing viewpoints. Furthermore, the proliferation of misinformation and fake news on social media platforms can exploit users' confirmation bias, thereby impacting the credibility and trustworthiness of VIs as sources of information. Consequently, understanding and mitigating confirmation bias is essential to ensure that VIs facilitate constructive dialogue and critical thinking among their audiences, thereby enhancing their effectiveness as marketing tools in SNS.

This phenomenon poses challenges to societal discourse and democracy, as it undermines the ability to engage in constructive dialogue and critical thinking (Lewandowsky et al., 2012). Nonetheless, developing media literacy appears as a critical aspect in mitigating the impact of confirmation bias, which correlates with a greater vulnerability to disinformation dissemination (Kalorth & Verma, 2018).

Next, to identify a means of interrupting confirmation bias, people tend to create cognitive dissonance (Chipidza & Yan, 2022). This approach involves introducing conflicting information or perspectives to challenge individuals' existing beliefs, prompting them to reevaluate their attitudes and behaviors. By instigating cognitive dissonance, individuals are encouraged to engage in critical reflection, fostering a more balanced and informed decision-making process.

2.3 Theory of Cognitive Dissonance

Cognitive dissonance, the discomfort experienced when holding conflicting beliefs or attitudes, has been extensively studied in psychology and has significant implications in the context of social media, SMIs, and VIs. Research suggests that social media platforms can exacerbate cognitive dissonance by exposing users to diverse viewpoints and conflicting information, leading to feelings of uncertainty and discomfort (Bail et al., 2018)Social media users may experience cognitive dissonance when encountering content that challenges their existing beliefs or values, prompting them to either reject opposing viewpoints or reassess their attitudes (Tandoc Jr., 2019). In the context of VIs cognitive dissonance may arise from the artificial nature of these personas and the discrepancy between their virtual

Email addresses: busyeuysal@solbridge.ac.kr (B.Uysal), restrella@solbridge.ac.kr (R.Estrella).

identities and the realities of human existence. Audiences may experience cognitive dissonance when engaging with VIs whose behavior or values diverge from their expectations of authentic human behavior. Additionally, the idealized representations of beauty and lifestyle promoted by VIs may contribute to cognitive dissonance by creating unattainable standards and aspirations among followers.

Next, to further challenge the user's established belief and perceptions regarding VIs, we explore the utilization of disclosure flags. A method of presenting the users with conflicting information about the identity of the VIs.

2.4 Disclosing Flags

Disclosing flags have emerged as a prevalent tool for users to report offensive content on various popular social media platforms. Therefore, it is used as a form of content moderation (Clune & McDaid, 2023). Their dual role effectively addresses the challenge of managing vast quantities of user-generated content and provides a justifiable basis for platform owners to remove content when necessary (Crawford & Gillespie, 2016).

From a theoretical perspective, research into user interpretations of content moderation reveals that disclosure flags are perceived as part of a larger, somewhat opaque system of platform moderation. These flags are frequently viewed as tools to balance the control exerted by platforms with the freedom of users, significantly influencing public perceptions of platform neutrality and fairness (Myers West, 2018). Although the intent behind using such flags is generally positive, inconsistent application can expose and even perpetuate broader societal biases and double standards, as discussed in the study "Double Standards in Social Media Content Moderation." Additionally, in scenarios involving prominent figures, the usage of disclosure flags can paradoxically increase user engagement through heightened visibility and public interest, despite their primary function as warnings or indicators of inaccuracies (Chipidza & Yan, 2022). In specific contexts such as health communities on social media, disclosure flags have been shown to positively influence user behavior and platform interaction, demonstrating their utility beyond general content moderation (Ysabel, 2018)

In case of VIs, employing disclosure flags is aimed at mitigating systematic risks. The implementation of these flags can be reinforced through updates to terms and conditions, along with robust enforcement mechanisms to ensure new requirements or restrictions are effectively integrated and adhered to (Mertens & Goetghebuer, 2024). This strategic use of disclosure flags seeks to enhance transparency and accountability, fostering a safer and more trustworthy digital environment. Moreover, this approach leverages the concept of cognitive dissonance, breaking users' cognitive biases and prompting a critical reassessment of the content they encounter, thereby enhancing informed decision-making.

3. Research Model and hypotheses

3.1 The impact of social information consumption on perceived trustworthiness

Heavy users are more likely to encounter diverse content, including both authentic and misleading information, which can shape their trust perceptions (Ao et al., 2023; Ryu & Han, 2021). Empirical evidence suggests that exposure to varied and interactive content enhances users' ability to evaluate the trustworthiness of the information they consume, fostering a more critical and trustful engagement with social media content (Lacap et al., 2023; Närvänen et al., 2020). Therefore, it is reasonable to infer that increased social media consumption is positively associated with perceived trustworthiness. Hence, hypothesis 1 is introduced as follows:

H1. Social information consumption on a social media platform is positively associated with perceived trustworthiness of VIs.

3.2 The Moderating Role of Knowledge Validation

Knowledge validation, defined as the process through which users confirm the accuracy of the information they consume, is pivotal in shaping user perceptions and trust. In this study, knowledge validation is operationalized through the

Email addresses: busyeuysal@solbridge.ac.kr (B.Uysal), restrella@solbridge.ac.kr (R.Estrella).

implementation of disclosure flags. These flags inform users that the content they are viewing is created by an artificial intelligent and promoted as a VIs rather than a SMIs. According to cognitive dissonance theory, such new information may cause users to experience psychological discomfort, prompting them to reassess their trust in the content and reevaluate their attitudes towards it (Festinger, 1962). Furthermore, confirmation bias suggests that users will likely interpret this information in a manner consistent with their preexisting beliefs about AI-generated content, potentially further diminishing their trust (Nickerson, 1998), therefore hypothesis 2 can be derived as follows:

H2. Knowledge validation through disclosure flags moderates the relationship between social media consumption on Instagram and perceived trustworthiness, potentially decreasing trust when users are aware of the non-human nature of the content creator.

3.3 The impact of perceived trustworthiness on affective behavior

Users may experience confirmation bias and cognitive dissonance when they discover that a VI is not human, as indicated by a disclosure flag. According to cognitive dissonance theory, individuals who initially trusted the message may feel psychological discomfort upon learning this fact. This discomfort can lead them to reevaluate their attitudes and actions toward the influencer, negatively impacting their positive engagement and affective behavior toward the content and platform (Festinger, 1962). Additionally, confirmation bias suggests that people tend to seek information that aligns with their existing beliefs. The disclosure flag confirming the non-human status of the influencer can reinforce users' skepticism towards AI-generated content, thereby diminishing their perceived trustworthiness of the VI. This amplified mistrust further reduces their emotional engagement and affective behavior (Metzger et al., 2018; Nickerson,

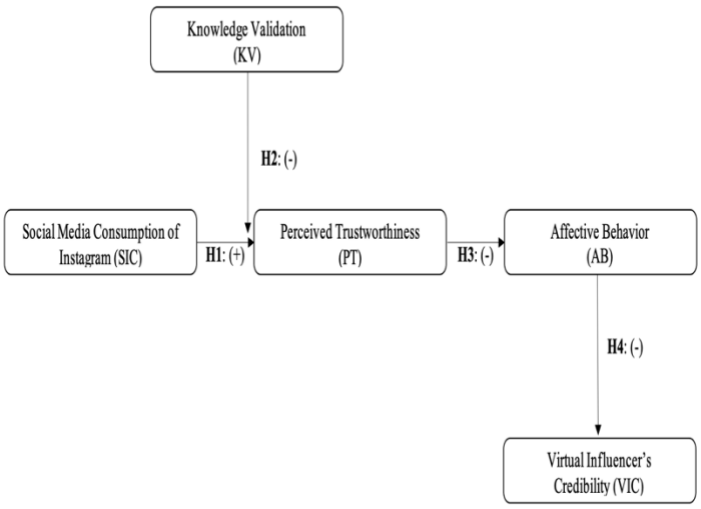

Fig. 1. Research Model

1998) thus:

H3. Perceived trustworthiness of the VI is negatively associated with affective behavior when users are aware of the influencer's non-human nature.

3.4 The impact of affective behavior on VI's credibility

Affective behavior, which encompasses the emotional responses and interactions users have with content, is crucial in shaping perceptions of the influencer's credibility. When users emotionally engage with content, such as liking, sharing, or commenting on posts, their perception of the influencer's credibility is likely enhanced (Fogg & Iizawa, 2008; Weiksner et al., 2008). Negative affective behavior, such as expressing distrust or disengaging with the content, can create a feedback loop that further undermines the influencer's credibility. It can be inferred that when users respond negatively, it signals to others that the influencer is not trustworthy, amplifying the skepticism towards the VI. Therefore:

H4. Affective behavior is negatively associated with the credibility of virtual influencers when users are aware of the influencer's non-human nature.

Email addresses: busyeuysal@solbridge.ac.kr (B.Uysal), restrella@solbridge.ac.kr (R.Estrella).

4.Data and Methodology

4.1 Survey design and data collection

4.1.1 Measurement Items

A questionnaire was created and disseminated to social media users as part of an instrument to gather data to test our hypothesis, drawing on earlier research on the behavior and behavior of social media users.

. As previously mentioned, to provide a better internal validity of our results, we investigated Instagram users, a dominating social media platform for photo sharing and one of the biggest breeding hubs for Vis, where Vis can be easily confused as read influencers (Xie-Carson et al., 2023). The study adapted the questionnaire items to meet the context of VIs research after gathering them from previous papers to ensure content validity. All the measurement items in the questionnaire were the seven-point Likert scale. The specific measurement scales are described in Appendix A.

Several questions were designed to grasp the demographic and socio-economic characteristics of respondents such as (1) gender, (2) age, (3) nationality, (4) level of education, (5) occupation (6) number of followers on Instagram, (7) number of accounts following on Instagram, (8) length of Instagram platform usage (9) Instagram profile anonymity status.

4.1.2 Data Collection

The study employed Qualtrics survey design software to structure the questionnaire. Two filtering questions were incorporated into the questionnaire to validate participants' status as Instagram users: (1) "Which of the following features is NOT available on Instagram?" and (2) "Which icon is typically used to 'like' a post on Instagram?". The survey repeats itself after disclosure flags are revealed to the participants. This acts as a form of knowledge validation in the survey that allows us to measure the effect of disclosing flags within individuals who took part in the survey.

The survey was distributed through dedicated, newly created profiles on the social media platforms Instagram and Reddit. These platforms were chosen for their extensive user bases and diverse demographics, enabling the recruitment of a varied sample from the Instagram user community.

At the end of the survey, a total of 202 participants completed the questionnaire. However, 11 responses were excluded due to failure to respond to both filtering questions. Additionally, upon review, 2 respondents were removed from the dataset as their repeated selection of identical answers suggested a lack of attention to the questionnaire. Consequently, 189 valid responses were included for analysis. Descriptive statistics pertaining to these 189 respondents are presented in Table 1.

4.2 Research Methodology

The versatility and applicability of partial least squares structural equation modeling (PLS-SEM) across various research contexts have resulted in a notable increase in its adoption among academics in recent years (Alsaad et al., 2018; Dash & Paul, 2021; Hair et al., 2011; Kurtaliqi et al., 2024). When modeling latent constructs amidst non-normal data distributions, PLS-SEM offers a robust analytical technique that imposes minimal constraints on measurement scales (Tenenhaus et al., 2005).

In executing PLS-SEM, the study prioritized ensuring the reliability and validity of the measurement model. This entailed employing an iterative application of ordinary least squares regression to derive outer weights, loadings, and structural relationships for both latent and manifest variables. Additionally, the study utilized bootstrap resampling to assess the statistical significance of structural paths.

5. Empirical Results

An exploratory factor analysis was carried out, and the findings are presented in Table 2. Each item's loading value on its corresponding latent variable, as shown in Table 2, exceeds all of its cross-loadings, indicating that the convergent and discriminant validity conditions were satisfied (Hair et al., 2011). After flagging, the discriminant validity remained intact, as each flagged item's loading on its respective construct continued to be higher than its cross-loadings with other constructs.

Before flagging, all constructs demonstrated strong internal consistency, as indicated by

Email addresses: busyeuysal@solbridge.ac.kr (B.Uysal), restrella@solbridge.ac.kr (R.Estrella).

Cronbach's Alpha (CA) and Composite Reliability (CR) values exceeding 0.7. Even after the flagging process, the reliability measures remained robust, with CA and CR values still above 0.7 for all constructs (Sarstedt et al., 2016). The average variance extracted (AVE) values were consistently above 0.5, confirming convergent validity both before and after the flagging process. This confirms that convergent validity was reliably achieved (Fornell & Larcker, 1981).

To evaluate discriminant validity, three types of statistical indicators were taken into consideration: (1) Cross loading, (2) Fornell-Larcker criterion, and (3) Heterotrait-Monotrait ratio (HTMT) (Sarstedt et al., 2016).

Table 4

Discriminant validity: Fornell-Larcker criterion.

Before Flag

	AB	C	PT	SIC
AB	0.802			
C	0.525	0.793		
PT	0.696	0.498	0.818	
SIC	0.393	0.179	0.342	0.795

After Flag

	AB_Flag	C_Flag	PT_Flag	SIC
AB_Flag	0.848			
C_Flag	0.560	0.806		
PT_Flag	0.743	0.472	0.863	
SIC	0.278	0.195	0.331	0.794

To determine discriminant validity, a matrix with cross-loading values was first examined (Gefen & Straub, 2005).

Each item's outer loading on its own construct had to be higher than its cross-loadings on other constructs. The cross-loadings indicated that each item's loading on the associated latent variable was greater than its loadings on other variables before flagging. Table 2 illustrates how the discriminant validity remained preserved even after flagging.

The second method used to assess discriminant validity was the Fornell-Larcker criterion. According to this criterion, the square root of the AVE for each latent variable must be larger than the highest correlation it has with any other latent variable (Fornell & Larcker, 1981). This method is known for being more conservative when evaluating discriminant validity (Sarstedt et al., 2016). Before flagging, each construct's square root of the AVE was larger than its correlations with other constructs. This trend continued after flagging, as evidenced by values such as AB_Flag (0.848), which maintained greater discriminant validity compared to its correlations with C_Flag (0.560) and other constructs.

Lastly, the HTMT ratio was applied to further evaluate discriminant validity. All HTMT values should be less than 0.85 (Henseler et al., 2015; Kline, 2023). Before flagging, all HTMT values were below this threshold, demonstrating strong discriminant validity. After flagging, HTMT values remained below 0.85, indicating that discriminant validity was preserved.

Table 5

Discriminant validity: HTMT.

Before Flag

	AB	C	PT	SIC
AB				
C	0.612			
PT	0.843	0.570		
SIC	0.489	0.224	0.414	

After Flag

	AB_Flag	C_Flag	PT_Flag	SIC
AB_Flag				
C_Flag	0.611			
PT_Flag	0.843	0.497		
SIC	0.324	0.216	0.384	

Considering these criteria, the research model satisfied all thresholds and conditions for discriminant validity both before and after flagging adjustments, ensuring a sufficient level of discriminant validity was secured.

After validating the measurement model, the research proceeded to estimate the structural model, which specifies the relationships between latent variables. Figures 2 and 3 illustrate the path

Email addresses: busyeuysal@solbridge.ac.kr (B.Uysal), restrella@solbridge.ac.kr (R.Estrella).

coefficients for the endogenous latent variables along with the R-squares. The initial analysis, depicted in Figure 2, examines the relationships between the variables before the introduction of the flag. The results reveal a significant association between Social Media Consumption of Instagram (SIC) and Perceived Trustworthiness (PT) ($\beta = 0.342$, $p < 0.001$). Furthermore, there is a significant relationship between PT and Affective Behavior (AB) ($\beta = 0.696$, $p < 0.001$). The association between AB and Virtual Influencer's Credibility (VIC) is also significant ($\beta = 0.525$, $p < 0.001$).

The subsequent analysis, illustrated in Figure 3, investigates the relationships after the flag was introduced to users. The results indicate a significant association between SIC and PT ($\beta = 0.331$, $p < 0.001$). Moreover, PT is significantly associated with AB ($\beta = 0.743$, $p < 0.001$). There is also a significant relationship between AB and VIC ($\beta = 0.560$, $p < 0.001$).

The introduction of the flag moderates the relationship between social media consumption on Instagram and perceived trustworthiness, as reflected in the slight decrease in the path coefficient (H2). Interestingly, however, the hypotheses suggesting negative associations between perceived trustworthiness and affective behavior (H3) and between affective behavior and VIC (H4) are not supported, as the associations remain positive even after the flag introduction. These findings indicate that while the flag influences perceived trustworthiness, it does not negatively impact affective behavior or the VIC.

6.Discussion and Conclusions

6.1. Theoretical Contributions

By integrating cognitive dissonance theory into the novel context of social media platforms, the paper illustrates that Instagram significantly enhances users' perceived trustworthiness of virtual influencers. This finding broadens our understanding of the dynamics of trust in digital environments, previously focused on privacy concerns (Jiang et al., 2013) and emotional responses (Krasnova et al., 2015). The study contributes to this discourse by introducing the moderating role of knowledge validation through disclosure flags, revealing that awareness of the non-human nature of content creators can moderate trust levels. This moderation effect aligns with cognitive dissonance theory, indicating that users experience a shift in trust when confronted with the reality of virtual influencers.

Moreover, our findings challenge existing assumptions about the impact of perceived trustworthiness on affective behavior. Contrary to expectations, it is observed that trust in virtual influencers can lead to positive affective behaviors, even when users are aware of the influencers' virtual nature. This outcome can be attributed to the selection of a highly realistic VI for the participants, which contrasts with other virtual influencers active on Instagram. The enhanced perception of humanity, combined with perceived trustworthiness, likely influenced the positive affective behavior of participants despite their awareness of the influencers' artificial nature (Cascio Rizzo et al., 2023)`. This indicates that realism and trust are critical factors in the effectiveness of virtual influencers. This insight contributes to the broader discourse on human-computer interaction and virtual personas in digital marketing, suggesting that the emotional engagement elicited by virtual influencers plays a crucial role in their perceived credibility.

6.2. Practical implications

The identified moderating effect of disclosure flags has significant implications for transparency in digital marketing. The findings suggest that clear disclosure of the virtual nature of influencers can influence user trust. As such, marketers and platform developers should implement transparent disclosure practices, labeling VIs and educating users about their nature to maintain an environment of honesty and reliability.

Additionally, the revelation that affective behavior positively influences VIs' credibility suggests that emotional engagement is critical for building and maintaining credibility. Marketers should design campaigns that not only inform but also emotionally resonate with their audience. This emotional connection can be achieved through storytelling, personalized content, and interactive

Email addresses: busyeuysal@solbridge.ac.kr (B.Uysal), restrella@solbridge.ac.kr (R.Estrella).

experiences, making the audience feel more connected to VIs.

Furthermore, the study highlights the importance of individual characteristics in moderating users' responses to disclosure flags. This insight offers a new perspective on the role of individual differences in the perception and behavior toward VIs, providing a direction for future research. Prior studies often used laboratory settings, which might overlook these individual traits. The findings suggest that real-world studies should consider these characteristics to better understand the dynamics of user engagement with VIs.

6.3Limitations and future research

While this study has its merits, there are some limitations that can be addressed in future research. First, Instagram was selected as the focal platform in this study, but there are other types of SNS, such as TikTok, Facebook, and Xiaohongshu. Second, as technology advances and VIs expand to other platforms, a more generalizable set of implications can be derived when further analysis is conducted on various kinds of platforms. Additionally, the sample size and demographic distribution may not be representative of the broader population, potentially limiting the generalizability of the findings. Longitudinal research designs could provide deeper insights into how user perceptions and behaviors toward virtual influencers evolve over time. Expanding the sample to include more diverse demographic groups could enhance the generalizability of the results.

Future work could delve into the territory of deepfakes as part of VIs, involving the disclosure of artificially generated or manipulated content, necessitating a deeper study into the ethics of VIs. As AI-generated content evolves and new forms of content appear, it is important to consider content moderation measures such as disclosure flags.

This research is both pertinent and timely given the ongoing proliferation of VIs. Despite current practices wherein companies manage and curate content on VI accounts, the potential emergence of autonomous VIs looms on the horizon. Thus, this study explores the imminent challenges stemming from this prospective scenario, notably the potential confusion among users in distinguishing between human and virtual entities. Furthermore, ethical considerations are paramount, prompting the advocacy for transparency and disclosure practices concerning VIs across diverse social media platforms. Consequently, there is a pressing need for policy interventions to navigate this dynamic landscape responsibly and uphold the integrity of online interactions.

Email addresses: busyeuysal@solbridge.ac.kr (B.Uysal), restrella@solbridge.ac.kr (R.Estrella).

Email addresses: busyeuysal@solbridge.ac.kr (B.Uysal), restrella@solbridge.ac.kr (R.Estrella).

Appendix A. Questionnaires

Table 1
Descriptive statistics.

Item		Frequency	Percentage (%)
Net Sample Size		189	100
Gender	Female	121	64.02
	Male	68	35.98
Age	<18	27	14.29
	19-24	133	70.37
	25-30	22	11.64
	31-35	2	1.06
	36-40	4	2.12
	>40	1	0.53
Nationality (passport holder)	North America	6	3.17
	South America	2	1.06
	Australia	2	1.06
	Europe	73	38.62
	Asia	95	50.26
	Africa	11	5.82
Education	Completed high school	63	33.33
	Completed/pursuing bachelor's	93	49.21
	Completed/pursuing master's or higher	20	10.58
	Completed/pursuing professional/vocational sch	13	6.88
Occupation	Employed full-time	24	12.7
	Employed part-time	9	4.76
	Self-employed	4	2.12
	Student	146	77.25
	Unemployed	6	3.17
Number of followers on Instagram	<200	57	30.16
	201-400	49	25.93
	401-600	28	14.81
	601-800	27	14.29
	801-1000	18	9.52
	>1000	12	6.35
Number of accounts following on Instagra	<200	53	28.04
	201-400	53	28.04
	401-600	31	16.4
	601-800	31	16.4
	801-1000	13	6.88
	>1001	8	4.23
Average use frequency of Instagram	Up to 30 minutes per day	32	16.93
	31 minutes – 1 hour per day	51	26.98
	1-3 hours per day	91	48.15
	More than 3 hours per day	15	7.94
Instagram Profile Status	Public	93	49.21
	Private	96	50.79

Email addresses: busyeuysal@solbridge.ac.kr (B.Uysal), restrella@solbridge.ac.kr (R.Estrella).

Appendix B.

Table 2

Discriminant Validity: Cross Loadings.

Before Flag				
	AB	C	PT	SIC
AB1	**0.779**	0.433	0.466	0.317
AB2	**0.821**	0.454	0.625	0.264
AB3	**0.886**	0.454	0.583	0.341
AB4	**0.711**	0.333	0.545	0.348
C1	0.444	**0.821**	0.411	0.149
C2	0.515	**0.787**	0.520	0.166
C3	0.351	**0.814**	0.313	0.190
C4	0.282	**0.749**	0.251	0.031
PT1	0.588	0.428	**0.781**	0.229
PT2	0.539	0.411	**0.822**	0.286
PT3	0.594	0.401	**0.876**	0.332
PT4	0.554	0.390	**0.788**	0.267
SIC1	0.351	0.175	0.264	**0.771**
SIC2	0.287	0.057	0.247	**0.849**
SIC3	0.281	0.172	0.277	**0.772**
SIC4	0.326	0.155	0.291	**0.786**

After Flag				
	AB_Flag	C_Flag	PT_Flag	SIC
AB_Flag 1	**0.842**	0.524	0.600	0.235
AB_Flag 2	**0.845**	0.508	0.639	0.196
AB_Flag 3	**0.921**	0.442	0.685	0.248
AB_Flag 4	**0.779**	0.421	0.592	0.267
C_Flag 1	0.487	**0.849**	0.334	0.156
C_Flag 2	0.582	**0.819**	0.577	0.226
C_Flag 3	0.312	**0.808**	0.285	0.152
C_Flag 4	0.306	**0.744**	0.194	0.038
PT_Flag 1	0.570	0.384	**0.833**	0.293
PT_Flag 2	0.652	0.398	**0.907**	0.289
PT_Flag 3	0.717	0.456	**0.857**	0.279
PT_Flag 4	0.611	0.382	**0.853**	0.282
SIC1	0.224	0.083	0.256	**0.767**
SIC2	0.144	0.062	0.207	**0.834**
SIC3	0.225	0.180	0.255	**0.766**
SIC4	0.263	0.253	0.309	**0.805**

Appendix C.

Table 3

Reliability and convergent validity.

Before Flag					
Construct	Outer Loading		CA	CR	AVE
Social Information Consumption (SIC)	SIC 1	0.771	0.805	0.805	0.632
	SIC 2	0.849			
	SIC 3	0.772			
	SIC 4	0.786			
Pereived Trust (PT)	PT 1	0.781	0.834	0.836	0.668
	PT 2	0.822			
	PT 3	0.876			
	PT 4	0.788			
Affective Behaviour (AB)	AB1	0.779	0.812	0.822	0.643
	AB2	0.821			
	AB3	0.886			
	AB4	0.711			
Credibility	C1	0.821	0.810	0.831	0.629
	C2	0.787			
	C3	0.814			
	C4	0.749			

After Flag					
Construct	Outer Loading		CA	CR	AVE
Social Information Consumption (SIC)	SIC 1	0.767	0.805	0.812	0.630
	SIC 2	0.834			
	SIC 3	0.766			
	SIC 4	0.805			
Pereived Trust (PT)	PT 1	0.833	0.885	0.889	0.744
	PT 2	0.907			
	PT 3	0.857			
	PT 4	0.853			
Affective Behaviour (AB)	AB1	0.842	0.868	0.872	0.719
	AB2	0.845			
	AB3	0.921			
	AB4	0.779			
Credibility	C1	0.849	0.828	0.872	0.649
	C2	0.819			
	C3	0.808			
	C4	0.744			

CA: Cronbach's Alpha, CR: Composite Reliability.

AVE: Average Variance Extracted.

Email addresses: busyeuysal@solbridge.ac.kr (B.Uysal), restrella@solbridge.ac.kr (R.Estrella).

Appendix D:

Table A.6

Measurement Items

Constructs	Item No.	Measurement Items	Reference
Social Information Consumption on Instagram	How often do you? (1= Never,7=Very often/a day)		
	AP1	Post photos	(Koroleva et al., 2011)
	AP2	Post stories	
	AP3	Share your thoughts and feelings on Instagram	
	AP4	Share somethings you are interested in	
Perceived Trustworthiness	PT1	The influencers can be relied upon on his/her content.	(Abou Ali et al., 2021)
	PT2	I believe what these influencers say and that he/she would not try to take advantage of their followers.	(Kim & Kim, 2021)
	PT3	These influencers are straightforward and transparent even though his/her self-interests are involved	
	PT 4	These influencers would not tell a lie even if he/she could profit from it	
Affective Behavior	AB1	I would recommend the influencers' accounts to strangers	(Ohanian, 1990)
	AB2	I would say positive things about the influencers' accounts to strangers	
	AB3	I would be likely to recommend the influencers to friends and relatives	
	AB4	I rarely passed up the chance to introduce others to these influences.	
Credibility	C1	These influencers remind me of someone who is competent and know what they are doing	(Erdem et al., 2004)
	C2	These influencers are honest	
	C3	These influencers are experts in their field	
	C4	These influencers are experienced	
Perceived Trustworthiness of VI	PT_Flag 1	The influencers can be relied upon on his/her content.	(Abou Ali et al., 2021)
	PT_Flag 2	I believe what these influencers say and that he/she would not try to take advantage of their followers.	(Kim & Kim, 2021)
	PT_Flag 3	These influencers are straightforward and transparent even though his/her self-interests are involved	
	PT_Flag 4	These influencers would not tell a lie even if he/she could profit from it	
Affective Behavior of VI	AB_Flag 1	I would recommend the influencers' accounts to strangers	(Ohanian, 1990)
	AB_Flag 2	I would say positive things about the influencers' accounts to strangers	
	AB_Flag 3	I would be likely to recommend the influencers to friends and relatives	
	AB_Flag 4	I rarely passed up the chance to introduce others to these influences.	
Credibility of VI	C_Flag 1	These influencers remind me of someone who is competent and know what they are doing	(Erdem et al., 2004)
	C_Flag 2	These influencers are honest	
	C_Flag 3	These influencers are experts in their field	
	C_Flag 4	These influencers are experienced	

Email addresses: busyeuysal@solbridge.ac.kr (B.Uysal), restrella@solbridge.ac.kr (R.Estrella).

Appendix E:

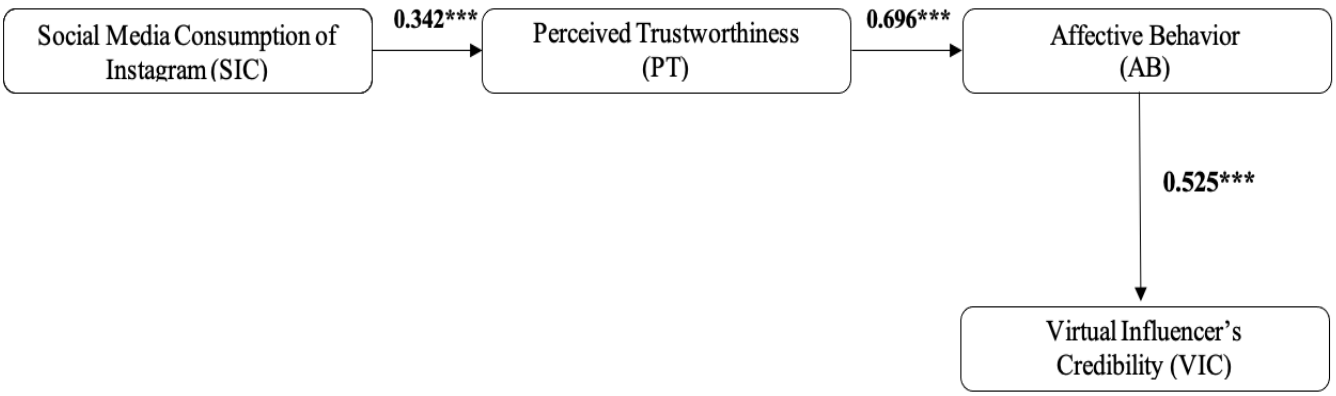

Fig. 2. Analysis results (structural model for before flagging).

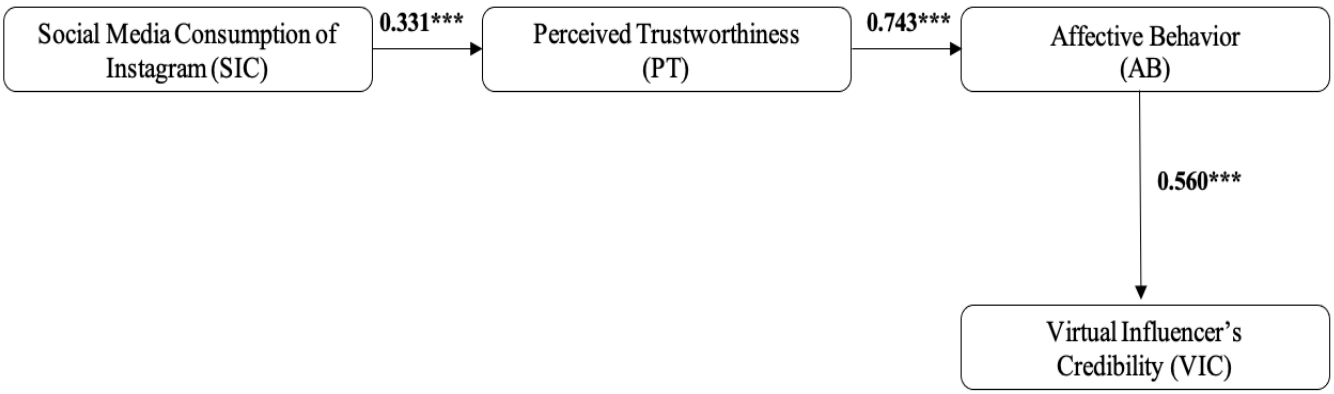

Fig. 3. Analysis results (structural model for after flagging).

Email addresses: busyeuysal@solbridge.ac.kr (B.Uysal), restrella@solbridge.ac.kr (R.Estrella).

References

Almeida, M. I. S. de, Coelho, R. L. F., Camilo-Junior, C. G., & Godoy, R. M. F. de. (2018). Quem Lidera sua Opinião? Influência dos Formadores de Opinião Digitais no Engajamento. *Revista de Administração Contemporânea*, *22*, 115–137. https://doi.org/10.1590/1982-7849rac2018170028

Alsaad, A., Taamneh, A., & Al-Jedaiah, M. N. (2018). Does social media increase racist behavior? An examination of confirmation bias theory. *Technology in Society*, *55*, 41–46. https://doi.org/10.1016/j.techsoc.2018.06.002

Antunes, A. C. (2022). The Role of Social Media Influencers on the Consumer Decision-Making Process. In *Research Anthology on Social Media Advertising and Building Consumer Relationships* (pp. 1420–1436). IGI Global. https://doi.org/10.4018/978-1-6684-6287-4.ch076

Ao, L., Bansal, R., Pruthi, N., & Khaskheli, M. B. (2023). Impact of Social Media Influencers on Customer Engagement and Purchase Intention: A Meta-Analysis. *Sustainability*, *15*(3), Article 3. https://doi.org/10.3390/su15032744

Arsenyan, J., & Mirowska, A. (2021). Almost human? A comparative case study on the social media presence of virtual influencers. *International Journal of Human-Computer Studies*, *155*, 102694. https://doi.org/10.1016/j.ijhcs.2021.102694

Bail, C. A., Argyle, L. P., Brown, T. W., & Volfovsky, A. (2018). *Exposure to opposing views on social media can increase political polarization | PNAS*. https://www.pnas.org/doi/abs/10.1073/pnas.1804840115

Bessi, A. (2016). Personality traits and echo chambers on facebook. *Computers in Human Behavior*, *65*, 319–324. https://doi.org/10.1016/j.chb.2016.08.016

Bringe, A. (2022). *The Rise Of Virtual Influencers And What It Means For Brands*. https://www.forbes.com/sites/forbescommunicationscouncil/2022/10/18/the-rise-of-virtual-influencers-and-what-it-means-for-brands/?sh=2e1014e76b56

Casaló, L. V., Flavián, C., & Ibáñez-Sánchez, S. (2020). Influencers on Instagram: Antecedents and consequences of opinion leadership. *Journal of Business Research*, *117*, 510–519. https://doi.org/10.1016/j.jbusres.2018.07.005

Cascio Rizzo, G. L., Berger, J. A., & Villarroel Ordenes, F. (2023). What Drives Virtual Influencer's Impact? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4329150

Chipidza, W., & Yan, J. (Kevin). (2022). The effectiveness of flagging content belonging to prominent individuals: The case of Donald Trump on Twitter. *Journal of the Association for Information Science and Technology*, *73*(11), 1641–1658. https://doi.org/10.1002/asi.24705

Choudhry, A., Han, J., Xu, X., & Huang, Y. (2022). "I Felt a Little Crazy Following a 'Doll'": Investigating Real Influence of Virtual Influencers on Their Followers. *Proceedings of the ACM on Human-Computer Interaction*, *6*(GROUP), 43:1-43:28. https://doi.org/10.1145/3492862

Chow, Y. (2023, February 1). Fad or future? Meet the virtual influencers taking over social media. *Stryve Digital Marketing*. https://www.stryvemarketing.com/blog/virtual-influencers/

Clune, C., & McDaid, E. (2023). Content moderation on social media: Constructing accountability in the digital space. *Accounting, Auditing & Accountability Journal*, *37*(1), 257–279. https://doi.org/10.1108/AAAJ-11-2022-6119

Conti, M., Gathani, J., & Tricomi, P. P. (2022). Virtual Influencers in Online Social Media. *IEEE Communications Magazine*, *60*(8), 86–91. https://doi.org/10.1109/MCOM.001.2100786

Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, *18*(3), 410–428. https://doi.org/10.1177/1461444814543163

Dash, G., & Paul, J. (2021). CB-SEM vs PLS-SEM methods for research in social sciences and technology forecasting. *Technological Forecasting and Social Change*, *173*, 121092. https://doi.org/10.1016/j.techfore.2021.121092

Email addresses: busyeuysal@solbridge.ac.kr (B.Uysal), restrella@solbridge.ac.kr (R.Estrella).

de Brito Silva, M. J., de Oliveira Ramos Delfino, L., Alves Cerqueira, K., & de Oliveira Campos, P. (2022). Avatar marketing: A study on the engagement and authenticity of virtual influencers on Instagram. *Social Network Analysis and Mining*, *12*(1), 130. https://doi.org/10.1007/s13278-022-00966-w

De Cicco, R., Iacobucci, S., Cannito, L., Onesti, G., Ceccato, I., & Palumbo, R. (2024). Virtual vs. human influencer: Effects on users' perceptions and brand outcomes. *Technology in Society*, *77*, 102488. https://doi.org/10.1016/j.techsoc.2024.102488

Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology*, *63*(4), 568–584. https://doi.org/10.1037/0022-3514.63.4.568

Djafarova, E., & Trofimenko, O. (2019). 'Instafamous' – credibility and self-presentation of micro-celebrities on social media. *Information, Communication & Society*, *22*(10), 1432–1446. https://doi.org/10.1080/1369118X.2018.1438491

Drenten, J., & Brooks, G. (2020). Celebrity 2.0: Lil Miquela and the rise of a virtual star system. *Feminist Media Studies*, *20*(8), 1319–1323. https://doi.org/10.1080/14680777.2020.1830927

Festinger, L. (1962). Cognitive Dissonance. *Scientific American*, *207*(4), 93–106.

Fogg, B. J., & Iizawa, D. (2008). Online Persuasion in Facebook and Mixi: A Cross-Cultural Comparison. In H. Oinas-Kukkonen, P. Hasle, M. Harjumaa, K. Segerståhl, & P. Øhrstrøm (Eds.), *Persuasive Technology* (pp. 35–46). Springer. https://doi.org/10.1007/978-3-540-68504-3_4

Fornell, C., & Larcker, D. F. (1981). Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. *Journal of Marketing Research*, *18*(1), 39–50. https://doi.org/10.1177/002224378101800104

Franke, C., Groeppel-Klein, A., & Müller, K. (2023). Consumers' Responses to Virtual Influencers as Advertising Endorsers: Novel and Effective or Uncanny and Deceiving? *Journal of Advertising*, *52*(4), 523–539. https://doi.org/10.1080/00913367.2022.2154721

Freberg, K., Graham, K., McGaughey, K., & Freberg, L. A. (2011). Who are the social media influencers? A study of public perceptions of personality. *Public Relations Review*, *37*(1), 90–92. https://doi.org/10.1016/j.pubrev.2010.11.001

Gefen, D., & Straub, D. (2005). A Practical Guide To Factorial Validity Using PLS-Graph: Tutorial And Annotated Example. *Communications of the Association for Information Systems*, *16*. https://doi.org/10.17705/1CAIS.01605

Geschke, D., Lorenz, J., & Holtz, P. (2019). The triple-filter bubble: Using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers. *British Journal of Social Psychology*, *58*(1), 129–149. https://doi.org/10.1111/bjso.12286

Geyser, W. (2020, February 18). *The State of Influencer Marketing 2020: Benchmark Report*. Influencer Marketing Hub. https://influencermarketinghub.com/influencer-marketing-benchmark-report/

Ghani, A. N. H. A., & Rahmat, H. (2023). Confirmation Bias in Our Opinions on Social Media: A Qualitative Approach. *Journal of Communication, Language and Culture*, *3*(1), Article 1. https://doi.org/10.33093/jclc.2023.3.1.4

Ha, L., & Yang, Y. (2023). Research about persuasive effects of social media influencers as online opinion leaders 1990-2020: A review. *International Journal of Internet Marketing and Advertising*, *18*(2–3), 220–241. https://doi.org/10.1504/IJIMA.2023.129661

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a Silver Bullet. *Journal of Marketing Theory and Practice*, *19*(2), 139–152. https://doi.org/10.2753/MTP1069-6679190202

Ham, J., Li, S., Shah, P., & Eastin, M. S. (2023). The "Mixed" Reality of Virtual Brand Endorsers: Understanding the Effect of Brand Engagement and Social Cues on Technological Perceptions and Advertising Effectiveness. *Journal of Interactive Advertising*, *23*(2), 98–113.

Email addresses: busyeuysal@solbridge.ac.kr (B.Uysal), restrella@solbridge.ac.kr (R.Estrella).

https://doi.org/10.1080/15252019.2023.2185557

Henseler, J., Ringle M., C., & Sarstedt, M. (2015). *A new criterion for assessing discriminant validity in variance-based structural equation modeling | Journal of the Academy of Marketing Science*. https://link.springer.com/article/10.1007/s11747-014-0403-8

Jang, H., & Yoh, E. (2020). Perceptions of male and female consumers in their 20s and 30s on the 3D virtual influencer. *The Research Journal of the Costume Culture*, *28*(4), 446–462. https://doi.org/10.29049/rjcc.2020.28.4.446

Jiang, Z., Heng, C. S., & C., B. (2013). *Research Note—Privacy Concerns and Privacy-Protective Behavior in Synchronous Online Social Interactions | Information Systems Research*. https://pubsonline.informs.org/doi/abs/10.1287/isre.1120.0441

Junkie, A. A. (2023, October 12). BMW Launches "Make it Real" Campaign With Virtual Creator Lil Miquela. *Branding in Asia*. https://www.brandinginasia.com/bmw-launches-make-it-real-campaign-with-virtual-creator-lil-miquela/

Kalorth, N., & Verma, M. (2018). Anatomy of Fake News: On (Mis)information and Belief in the Age of Social Media. *Journal of Content, Community and Communication*, *4*(8), 9–14. https://doi.org/10.31620/JCCC.12.18/03

Khamis, S., Ang, L., & Welling, R. (2017). Self-branding, 'micro-celebrity' and the rise of Social Media Influencers. *Celebrity Studies*, *8*(2), 191–208. https://doi.org/10.1080/19392397.2016.1218292

Kim, D., & Wang, Z. (2023). The ethics of virtuality: Navigating the complexities of human-like virtual influencers in the social media marketing realm. *Frontiers in Communication*, *8*. https://doi.org/10.3389/fcomm.2023.1205610

Kim, H., & Park, M. (2023). Virtual influencers' attractiveness effect on purchase intention: A moderated mediation model of the Product–Endorser fit with the brand. *Computers in Human Behavior*, *143*, 107703. https://doi.org/10.1016/j.chb.2023.107703

Kline, R. B. (2023). *Principles and Practice of Structural Equation Modeling*. Guilford Publications.

Koles, B., Audrezet, A., Moulard, J. G., Ameen, N., & McKenna, B. (2024). The authentic virtual influencer: Authenticity manifestations in the metaverse. *Journal of Business Research*, *170*, 114325. https://doi.org/10.1016/j.jbusres.2023.114325

Krasnova, H., Widjaja, T., Buxmann, P., Wenninger, H., & Benbasat, I. (2015). Research Note—Why Following Friends Can Hurt You: An Exploratory Investigation of the Effects of Envy on Social Networking Sites among College-Age Users. *Information Systems Research*, *26*(3), 585–605. https://doi.org/10.1287/isre.2015.0588

Kurtaliqi, F., Lancelot Miltgen, C., Viglia, G., & Pantin-Sohier, G. (2024). Using advanced mixed methods approaches: Combining PLS-SEM and qualitative studies. *Journal of Business Research*, *172*, 114464. https://doi.org/10.1016/j.jbusres.2023.114464

Lacap, J. P. G., Cruz, M. R. M., Bayson, A. J., Molano, R., & Garcia, J. G. (2023). Parasocial relationships and social media interactions: Building brand credibility and loyalty. *Spanish Journal of Marketing - ESIC*, *28*(1), 77–97. https://doi.org/10.1108/SJME-09-2022-0190

Lanius, C., Weber, R., & MacKenzie, W. I. (2021). Use of bot and content flags to limit the spread of misinformation among social networks: A behavior and attitude survey. *Social Network Analysis and Mining*, *11*(1), 32. https://doi.org/10.1007/s13278-021-00739-x

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). *Misinformation and Its Correction: Continued Influence and Successful Debiasing*. https://journals.sagepub.com/doi/full/10.1177/1529100612451018

Lou, C., Kiew, S. T. J., Chen, T., Lee, T. Y. M., Ong, J. E. C., & Phua, Z. (2023). Authentically Fake? How Consumers Respond to the Influence of Virtual Influencers. *Journal of Advertising*, *52*(4), 540–557.

Email addresses: busyeuysal@solbridge.ac.kr (B.Uysal), restrella@solbridge.ac.kr (R.Estrella).

https://doi.org/10.1080/00913367.2022.2149641

Mertens, F., & Goetghebuer, J. (2024). *Virtual Reality, Real Responsibility: The Regulatory Landscape for Virtual Influencers* (SSRN Scholarly Paper 4718820). https://doi.org/10.2139/ssrn.4718820

Meta. (2024). Our Approach to Labeling AI-Generated Content and Manipulated Media. *Meta*. https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/

Metzger, A., Alvis, L. M., Oosterhoff, B., Babskie, E., Syvertsen, A., & Wray-Lake, L. (2018). The Intersection of Emotional and Sociocognitive Competencies with Civic Engagement in Middle Childhood and Adolescence. *Journal of Youth and Adolescence*, *47*(8), 1663–1683. https://doi.org/10.1007/s10964-018-0842-5

Modgil, S., Singh, R. K., Gupta, S., & Dennehy, D. (2024). A Confirmation Bias View on Social Media Induced Polarisation During Covid-19. *Information Systems Frontiers*, *26*(2), 417–441. https://doi.org/10.1007/s10796-021-10222-9

Moustakas, E., Lamba, N., Mahmoud, D., & Ranganathan, C. (2020). Blurring lines between fiction and reality: Perspectives of experts on marketing effectiveness of virtual influencers. *2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security)*, 1–6. https://doi.org/10.1109/CyberSecurity49315.2020.9138861

Muttamimah, L., & Irwansyah, I. (2023). Pemanfaatan Influencer Berbasis Virtual dalam Komunikasi Pemasaran. *WACANA: Jurnal Ilmiah Ilmu Komunikasi*, *22*(1), Article 1. https://doi.org/10.32509/wacana.v22i1.2322

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, *20*(11), 4366–4383. https://doi.org/10.1177/1461444818773059

Närvänen, E., Kirvesmies, T., & Kahri, E. (2020). Parasocial relationships of Generation Z consumers with social media influencers. In *Influencer Marketing*. Routledge.

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Parsani, P. (2023). *Case Study: The AI Behind Virtual Influencer Lil Miquela*. https://www.cut-the-saas.com/ai/the-ai-behind-virtual-influencer-lil-miquela

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

Phua, J., Jin, S. V., & Kim, J. (Jay). (2017). Gratifications of using Facebook, Twitter, Instagram, or Snapchat to follow brands: The moderating effect of social comparison, trust, tie strength, and network homophily on brand identification, brand engagement, brand commitment, and membership intention. *Telematics and Informatics*, *34*(1), 412–424. https://doi.org/10.1016/j.tele.2016.06.004

Premia, P. (2022). *Asia metaverse – the coming of age of virtual influencers*. Premia Partners. https://www.premia-partners.com/insight/asia-metaverse-the-coming-of-age-of-virtual-influencers

Rahkman Ardi, A. (2021). Determinant factors of partisans' confirmation bias in social media. *Humanitas Indonesian Psychological Journal*, *18*(1), Article 1.

Robinson, B. (2020). Towards an Ontology and Ethics of Virtual Influencers. *Australasian Journal of Information Systems*, *24*. https://doi.org/10.3127/ajis.v24i0.2807

Ryu, E. A., & Han, E. (2021). Social Media Influencer's Reputation: Developing and Validating a Multidimensional Scale. *Sustainability*, *13*(2), Article 2. https://doi.org/10.3390/su13020631

Sands, S., Ferraro, C., Demsar, V., & Chandler, G. (2022). False idols: Unpacking the opportunities and challenges of falsity in the context of virtual influencers. *Business Horizons*, *65*(6), 777–788.

Email addresses: busyeuysal@solbridge.ac.kr (B.Uysal), restrella@solbridge.ac.kr (R.Estrella).

https://doi.org/10.1016/j.bushor.2022.08.002

Sarstedt, M., Hair, J. F., Ringle, C. M., Thiele, K. O., & Gudergan, S. P. (2016). Estimation issues with PLS and CBSEM: Where the bias lies! *Journal of Business Research*, *69*(10), 3998–4010. https://doi.org/10.1016/j.jbusres.2016.06.007

Scherer, M. U. (2015). *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies* (SSRN Scholarly Paper 2609777). https://doi.org/10.2139/ssrn.2609777

Statista. (2024). *Influencer Advertising—Global | Statista Market Forecast*. Statista. https://www.statista.com/outlook/amo/advertising/influencer-advertising/worldwide

Taber, C. S., & Lodge, M. (2006). *Motivated Skepticism in the Evaluation of Political Beliefs—Taber—2006—American Journal of Political Science—Wiley Online Library*. https://onlinelibrary.wiley.com/doi/full/10.1111/j.1540-5907.2006.00214.x

Tandoc Jr., E. C. (2019). The facts of fake news: A research review. *Sociology Compass*, *13*(9), e12724. https://doi.org/10.1111/soc4.12724

Tenenhaus, M., Vinzi, V. E., Chatelin, Y.-M., & Lauro, C. (2005). PLS path modeling. *Computational Statistics & Data Analysis*, *48*(1), 159–205. https://doi.org/10.1016/j.csda.2004.03.005

TikTok. (2023). *New labels for disclosing AI-generated content*. Newsroom | TikTok. https://newsroom.tiktok.com/en-us/new-labels-for-disclosing-ai-generated-content

Time. (2018, June 28). *Meet This Year's 25 Most Influential People on the Internet*. TIME. https://time.com/5324130/most-influential-internet/

Um, N. (2023). Predictors Affecting Effects of Virtual Influencer Advertising among College Students. *Sustainability*, *15*(8), Article 8. https://doi.org/10.3390/su15086388

Weiksner, G. M., Fogg, B. J., & Liu, X. (2008). Six Patterns for Persuasion in Online Social Networks. In H. Oinas-Kukkonen, P. Hasle, M. Harjumaa, K. Segerståhl, & P. Øhrstrøm (Eds.), *Persuasive Technology* (pp. 151–163). Springer. https://doi.org/10.1007/978-3-540-68504-3_14

Xie-Carson, L., Benckendorff, P., & Hughes, K. (2023). Keep it #Unreal: Exploring Instagram Users' Engagement With Virtual Influencers in Tourism Contexts. *Journal of Hospitality & Tourism Research*, 10963480231180940. https://doi.org/10.1177/10963480231180940

Xin, B., Hao, Y., & Xie, L. (2024). Virtual influencers and corporate reputation: From marketing game to empirical analysis. *Journal of Research in Interactive Marketing*, *ahead-of-print*(ahead-of-print). https://doi.org/10.1108/JRIM-10-2023-0330

Ysabel, gerrard. (2018). *Beyond the hashtag: Circumventing content moderation on social media*. https://journals.sagepub.com/doi/full/10.1177/1461444818776611

Yu, J., Dickinger, A., So, K. K. F., & Egger, R. (2024). Artificial intelligence-generated virtual influencer: Examining the effects of emotional display on user engagement. *Journal of Retailing and Consumer Services*, *76*, 103560. https://doi.org/10.1016/j.jretconser.2023.103560

Zhao, H., Fu, S., & Chen, X. (2020). Promoting users' intention to share online health articles on social media: The role of confirmation bias. *Information Processing & Management*, *57*(6), 102354. https://doi.org/10.1016/j.ipm.2020.102354

Zhong, L. (2022). *Analyses of the Relationship between Virtual Influencers' Endorsements and Customer Brand Engagement in Social Media*. 37–41. https://doi.org/10.2991/aebmr.k.220404.007

Email addresses: busyeuysal@solbridge.ac.kr (B.Uysal), restrella@solbridge.ac.kr (R.Estrella).