VISUAL REVIEW | Vol. 17, No. 4, 2025 | ISSN 2695-9631
International Visual Culture Review / Revista Internacional de Cultura Visual
https://doi.org/10.62161/revvisual.v17.5864

MISINFORMATION DISGUISED AS HUMOUR: Deepfakes' Impact on Instagram Trust

FERNANDO FUENTE-ALBA CARIOLA[1]
ffuentealba@ucsc.cl

CLAUDIO TORRES ARAVENA[1]
ctorres@ucsc.cl

[1]Universidad Católica de la Santísima Concepción, Chile

KEYWORDS

Deepfakes
Artificial intelligence
Trust
Information
Instagram
Chile

ABSTRACT

This research analyses the impact of deepfakes, i.e. fake videos generated by artificial intelligence, on trust in Instagram in Concepción, Chile. In addition to analysing deepfakes, focus groups and questionnaires were conducted with over 100 people. Notably, 85.2% of participants reported having seen at least one deepfake video on Instagram. When they identify one, more than half of them laugh. 57.6% admitted to doubting its authenticity, and 81.4% indicated that deepfakes affect their trust in Instagram.

Received: 11/ 04 / 2025
Accepted: 05/ 07 / 2025

1. Introduction

The employment of artificial intelligence to generate synthetic videos, otherwise referred to as “deepfakes”, poses a substantial challenge to the integrity and reliability of information disseminated on social networks. It is imperative to comprehend the potentially deleterious consequences of such videos on society, particularly given that their unregulated propagation has the potential to erode public trust in institutions, social networks and traditional media (Garcia, 2021). The significance of the research problem concerning the impact of these fake videos on trust on Instagram is substantial within the discipline of communication and journalism. The term “deepfakes” is used to describe forged videos or images generated by artificial intelligence. It has been argued that deepfakes have the potential to disrupt public trust in social media and undermine the integrity of information. The challenge to truthfulness in the media and social networks is intensified by the rise of deepfakes, which constitute a novel form of misinformation capable of compromising trust in digital media. This study explores the potential of advanced manipulation techniques to compromise trust on visual social media platforms such as Instagram, a milieu already vulnerable to visual misinformation, and the subsequent impact on users' perceptions of the network.

As Molina (2024) observes, the capacity of artificial intelligence to generate realistic fake videos poses significant challenges to the integrity of information and public trust in media outlets. This is particularly salient given that the majority of individuals are unable to discern the falsity of the production, thereby increasing their vulnerability to misinformation (Ballesteros-Aguayo and Ruiz del Olmo, 2024). The term "deepfake" was first coined in 2017 on internet forums, specifically on Reddit, by a user who went by the name of "deepfakes". This user began to share videos of celebrities that had been created using artificial intelligence (AI) techniques, combining the word "deep learning" with "fake" (Chesney & Citron, 2019). The popularity of such videos can be attributed to the sophistication of the visual effects and the simplicity of their production. Nevertheless, it was in the political arena that deepfakes were established as a significant threat.

A notable instance of this phenomenon was the circulation of a manipulated video of Nancy Pelosi, the Speaker of the US House of Representatives, in 2019. In this video, Pelosi appeared to be speaking in a incoherent manner, which was swiftly disseminated via social media, giving rise to concerns regarding her state of mental health (Vaccari & Chadwick, 2020). While this video did not constitute a deepfake in the strict sense of the term (it had been modified by means of less sophisticated editing techniques), its circulation did demonstrate the potential impact of such content in a political context.

Since then, several instances of deepfakes have attracted global attention. A notable instance of this phenomenon occurred in 2020, when a fabricated video of former US President Barack Obama was created by filmmaker Jordan Peele, in collaboration with the news website BuzzFeed. The purpose of the video was to demonstrate the potential of this technology to disseminate disinformation (Ajder et al., 2019). In the video, Obama appeared to warn about the dangers of disinformation and fake news. However, the actual message was delivered by Peele, who used deepfake technology to synchronise Obama's lip movements and expressions with the filmmaker's words.

Conversely, Wardle and Derakhshan (2017) address the concept of disinformation as part of the broader phenomenon of information clutter. The present disorder has been shown to comprise three distinct types of information-related problems, namely: disinformation (where the intention is to mislead through the propagation of false information), misinformation (where the intention is to cause harm through the propagation of true information) and misinformation (where the intention is not to mislead through the propagation of false information). The term "deepfakes" has been identified as a form of disinformation, characterised by the creation of authentic-looking audio-visual material with the intent to deceive. The authors emphasise that this form of content has the potential to erode public trust in conventional media outlets, particularly when it is extensively disseminated through social networking platforms such as Instagram or Facebook, where the mechanisms for content verification are often limited.

Lazer et al. (2018) posit that the dissemination of false information, including the creation of manipulated audiovisual content, poses a significant threat to democratic institutions and the integrity of the digital society. It is asserted that social media platforms have a role to play in the rapid dissemination of manipulated content, a factor that is widely considered to exacerbate the situation.

The dissemination of false information and the manipulation of digital media through deepfake technology has been demonstrated to have a significant impact on societal polarisation. It has been observed that individuals often share content that aligns with their pre-existing beliefs, without undertaking the necessary steps to verify its veracity. Indeed, there are authors such as Ibrahim (2021) who undertake exhaustive analyses of the psychological effects of deepfakes on public perception.

It is contended that humans are predisposed to place trust in visual and auditory evidence, a susceptibility that renders deepfakes particularly perilous.

The manipulation of images and videos represents a significant attack on the primary means by which individuals construct their understanding of the world. In contrast to other forms of disinformation, deepfakes have been shown to have a particularly potent emotional impact, given their capacity to deceive both the intellect and the senses. This has the effect of complicating efforts to verify information, and of deepening distrust of media and official sources.

It is imperative that regulation and governance mechanisms ensure the ethical and responsible use of Artificial Intelligence (AI) in the media, taking into consideration the practical implications of its implementation. The development of AI-generated video constitutes a substantial technological advancement; however, it concomitantly gives rise to grave ethical and social concerns. According to Hueso (2022), the absence of regulatory frameworks and supervisory mechanisms in this domain engenders a permissive environment for the malevolent exploitation of deepfakes, facilitating the dissemination of manipulation and misinformation. The increasing sophistication of these videos poses a significant threat to the authenticity and veracity of information disseminated on social media platforms. Fletcher and Park (2017) posit that "trust in information channels is an essential component of a healthy democracy" (p. 3). The absence of suitable regulatory frameworks to address the creation and dissemination of these realisations poses a significant threat to the security of information and the trust placed in democratic processes. By providing empirical evidence and detailed analysis, this study can inform policy discussions and help develop measures to mitigate the negative effects of fake videos (Arteaga et al., 2023).

It is imperative to comprehend the impact of these audiovisual productions on the trust of their viewers to formulate effective strategies that safeguard the integrity of information and public trust. Indeed, as posited by authors such as Chesney and Citron (2019), "deepfakes have the potential to be employed for the dissemination of disinformation on a large scale, which could have a profoundly deleterious effect on public opinion and social cohesion" (p. 1754).

In the context of Chile and Latin America, where social media plays a pivotal role in shaping public opinion, the proliferation of disinformation through video content has the potential to erode citizens' trust in media institutions and impede their ability to make informed decisions. Furthermore, the capacity of artificial intelligence to generate content that is strikingly realistic gives rise to significant ethical considerations. As Vaccari and Chadwick (2020) observe, "the manipulation of images and videos by AI poses unique challenges for information ethics, as it can be difficult, even for experts, to distinguish between the real and the fake" (p. 6). This difficulty may be even more pronounced among the general population, who may not be familiar with the existence or capabilities of deepfakes, an issue that is clear in this research where more than 80% of respondents to the questionnaire are unable to accurately identify whether it is a fake video or not.

The overarching aim of this research is thus pertinent, namely, to analyse the impact of deepfakes generated by artificial intelligence on the trust of Instagram users in the city of Concepción, Chile. The objective will be addressed through the following specific objectives: firstly, identifying and analysing the most prevalent types of deepfakes generated by artificial intelligence in Instagram posts in Chile during 2024; secondly, interpreting the trust and perception of Instagram users regarding the exposure of such fake videos; and thirdly, categorising the expectations and experiences of Instagram users in Concepción, Chile, regarding

This is a term that has been coined to describe the practice of using artificial intelligence to create realistic fake videos, often with the intention of deceiving an audience. The term "deepfakes" was first popularised on the social media platform Instagram.

The Chilean media landscape is characterised by a diverse array of both traditional and digital media outlets, which collectively facilitate access to a plethora of data sources for analysis. It is evident that social media and online news platforms are characterised by a substantial abundance of data, rendering them conducive environments for the identification and study of deepfakes. Furthermore, collaboration with academic institutions and media organisations has been demonstrated to facilitate access to data and archives required for research (Hu et al., 2020).

The feasibility of this research is founded on the availability of technological and academic resources, access to data, applicability of appropriate methodologies, and potential institutional support. These conditions permit meaningful research on the impact of deepfakes on the veracity and trustworthiness of users of the social network Instagram through a case study in the city of Concepción, Chile.

This research makes a significant contribution to the field of digital communication and social media studies, given that news consumption and interaction with content on Instagram is growing. The assessment of the impact of deepfakes in this context is conducive to the generation of insights into the manner in which emerging technologies affect public trust, a central theme in contemporary media research.

2. Design and Method

2.1. Formal object of study

The present research seeks to analyse the impact of deepfakes on the trust of users of the social network Instagram, through a case study located in the city of Concepción, Chile. This approach is predicated on the premise that such videos, created by artificial intelligence, not only challenge the public's trust in the information shared on digital platforms, but also give rise to ethical dilemmas and transform communicative practices. As Zuboff (2019) observes, advancements in media manipulation technologies, such as deepfakes, necessitate critical reflection on the ethical responsibility of communicators, the impact of these technologies on the integrity of information, and the risk of misinformation in the media. The selection of cases will facilitate the identification of patterns, perceptions and attitudes towards manipulated content, thereby generating an understanding of its effects on the digital information ecosystem. This methodological approach facilitates not only a theoretical examination of the phenomenon, but also an empirical analysis of the specific dynamics among Instagram users and their interaction with deepfakes. This empirical analysis enables a more profound reflection on training in communication ethics and critical competence in the face of current technological challenges.

2.2. The nature of the research

The research method employed was a combination of qualitative and quantitative techniques (Sierra Bravo, 1992). This methodology combines the in-depth interpretation of social and cultural phenomena with the ability to measure and analyse general trends using quantitative data. From a qualitative standpoint, the objective of the present study is to explore how a select group of users perceive and react to content that has been manipulated through the use of deepfakes. This methodological approach facilitates the interpretation of the underlying social and contextual dynamics of an emerging phenomenon, thereby providing a more profound insight into the experiences of both individuals and groups. Sierra Bravo (1992) posits that qualitative methodologies are especially efficacious in research endeavours concerning intricate phenomena, such as the phenomenon of digital manipulation within social networks. In terms of the quantitative approach, tools such as closed-ended questionnaires are utilised to collect structured data on the frequency of exposure to these fake videos, levels of trust in the information consumed, and data regarding Instagram users' experiences with these artificial intelligence realisations. In the quantitative domain, a content analysis will be conducted on the most visited and virally disseminated deepfakes in Chile during 2024, with the objective of classifying and categorising them. This component facilitates the identification of patterns and relationships between key variables, thereby enabling more effective analysis by category. The employment of a mixed approach guarantees a methodological approach that facilitates in-depth exploration of the subjective perceptions of Instagram users, whilst concurrently enabling systematic analysis of the effects of deepfakes on the trustworthiness of visual content on the platform. In this framework, an analysis of selected cases will be used to understand how users interact with manipulated content and how it affects their trust in the information. This methodological approach guarantees an approach to the object of study, whilst simultaneously ensuring the necessary flexibility to capture the particularities of the phenomenon under investigation.

2.3. Hypothesis

The present study investigates the impact of deepfakes, generated by artificial intelligence, on the perception and trust of users in the veracity of information disseminated on the social network Instagram, with a particular focus on users residing in Concepción, Chile. This hypothesis posits that exposure to deepfakes impacts the perceived authenticity of information, thereby influencing trust in social networks, particularly in contexts where information accuracy is paramount, such as in the political arena (Martínez, 2023).

Indeed, the dissemination of deepfakes has the potential to compromise the reliability of information, thereby reflecting concerns regarding the influence of misinformation and manipulation on perceptions of reality (Ballesteros-Aguayo and Ruiz del Olmo, 2024).

2.4. Scope of study

The present study focuses on analysing the impact of deepfakes generated by artificial intelligence on the trust of users in Concepción, Chile. The present study specifically addresses the question of how deepfakes affect the trustworthiness of information disseminated by the social network Instagram, as well as the ethical and social implications derived from their use. The universe under investigation in this research study is consistent with the characteristics of a purposive sampling study, insofar as it seeks to understand an emerging phenomenon in a specific local context. This type of study can be framed within a mixed methodology, integrating qualitative and quantitative approaches to provide a more comprehensive and in-depth understanding of the phenomenon under study (Creswell, 2014).

2.5. Variables of analysis

The research considers people with varying levels of knowledge about deepfakes and digital manipulation, as well as a variety of ages, genders, interests and levels of digital literacy. The selection criteria are as follows: The geographical focus of the study is Gran Concepción; the demographic focus is as follows: This study incorporates a range of users, including those with extensive experience and potential comprehension of deepfakes, their impact on trust in digital content, and the thematic implications of this phenomenon. The sample is constrained to active Instagram users who demonstrate regular interaction on this social network, acknowledging its role as a primary platform for the dissemination and consumption of manipulated visual content. The present study concentrated on the Chilean city of Concepción. The deepfakes selected for documentary analysis and focus group discussion correspond to the eight videos made by artificial intelligence that have received the greatest number of views and "likes" on the Instagram network in Chile during 2024.

2.6. Research techniques

The methodology to be employed in this research comprises a combination of quantitative and qualitative techniques.

This paper sets out a methodological strategy that combines quantitative and qualitative techniques with a view to obtaining an in-depth understanding of the impact of deepfakes on the trust of users of the Instagram social network. To this end, a range of collection and analysis techniques will be employed.

2.6.1. Data Collection Techniques

The initial instrument employed was a semi-structured questionnaire, administered in digital format via the Google Forms platform. The questionnaire has been meticulously designed to capture both quantitative and qualitative data through a combination of closed and open-ended questions. The instrument's structure is organised into thematic sections that seek to investigate the daily use of Instagram by the participants, their exposure to content generated through deepfakes and the level of trust that they express towards the publications they find on this social network. The sample comprised 107 individuals who were active on Instagram, including 53 women, 50 men and four who did not disclose their gender preference. These individuals were selected through non-probabilistic convenience sampling, ensuring diversity in terms of age, gender and level of interaction with the platform. This methodological approach enabled a descriptive analysis of trends and patterns, while integrating qualitative responses into the thematic analysis to facilitate a more profound interpretation of the observed phenomena.

The second instrument employed was a documentary analysis, the focus of which was eight deepfake videos that had been selected on the basis of their relevance and circulation on Instagram. The videos encompass depictions of both Chilean and foreign individuals, thereby ensuring a diversity of styles and levels of manipulation. For this stage of the analysis, an evaluation matrix was applied. This was designed to evaluate aspects such as technical style (resolution, quality of manipulation), the type of alteration (visual, auditory or mixed) and the perceived purpose of each content (humorous, informative, uninformative or political). This analysis enabled us to ascertain how the intrinsic characteristics of the videos influence the perception of their authenticity and the emotional response of the users. The third instrument corresponds to a focus group composed of seven people living in the metropolitan area of Gran Concepción.

The selection of members was made with a view to achieving heterogeneity with regard to age, gender and experience of social network use. During the session, participants were shown the same series of deepfakes videos that had been selected for the documentary analysis. Following this, the participants were presented with a questionnaire intended to elicit their immediate and reflective perceptions of the videos. The questionnaire focused on the perceived authenticity of the videos, the emotional impact experienced by the participants, and the implications of the videos.

The identification of the phenomenon is characterised by an analysis of the level of trust placed in Instagram content. The focus group dynamics encompassed a moderated open discussion, which was recorded, transcribed and qualitatively analysed using thematic coding techniques.

3. Fieldwork and Analysis Techniques

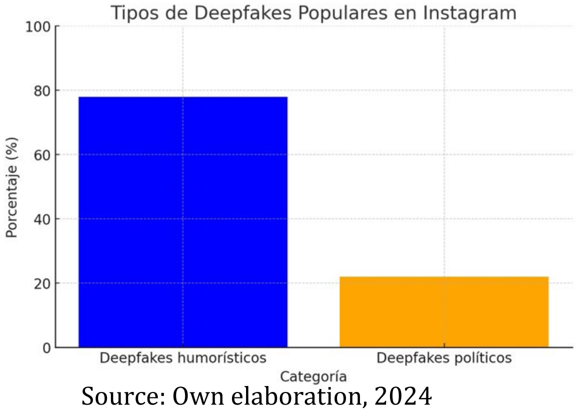
The integration of the data from these three instruments was achieved through methodological triangulation. This methodological triangulation approach enabled the synthesis of findings from three distinct sources: the questionnaire, the documentary analysis, and the focus group. This methodological triangulation approach was adopted to enhance the validity and reliability of the findings. The quantitative data obtained from the online questionnaire were analysed using descriptive statistical techniques, allowing the identification of general patterns in the sample. Conversely, the qualitative data, derived from both the documentary analysis and the focus group, underwent a thematic analysis with the objective of identifying key categories, relationships between concepts, and emergent meanings. This triangulation of data enabled the validation of the findings and the offer of a nuanced interpretation of the impact of deepfakes on the trust of Instagram users. This addressed the complexity of the phenomenon studied from different perspectives.

4. Results

The triangulation of the data obtained is intended to integrate the findings from three methodological tools: online questionnaire, focus group and analysis of the eight selected videos. This combination of methods allows us to corroborate the hypothesis regarding the negative impact of deepfakes on the perception of users in the Metropolitan Area of Greater Concepción regarding the veracity of information on Instagram. Furthermore, the specific objectives defined in the study are addressed through different categories:

The style of the deepfakes is characterised by both visual and narrative elements, with consideration given to factors such as technical and aesthetic quality. The utilisation of this category facilitated the identification of common patterns and characteristics within the videos that were analysed. With regard to this category, Figure 1 demonstrates that 78% of the identified deepfakes correspond to a humorous style and 22% to a political style, as determined by the selection patterns. It is noteworthy that the humorous videos focus on prominent figures, including the current US President Donald Trump, the actor Tom Cruise, and the Chilean footballer Alexis Sánchez. As has been demonstrated, the political deepfakes have been observed to focus on figures such as the Chilean President, Gabriel Boric, and the Argentinean President, Javier Milei.

Figure 1 Types of Deepfakes

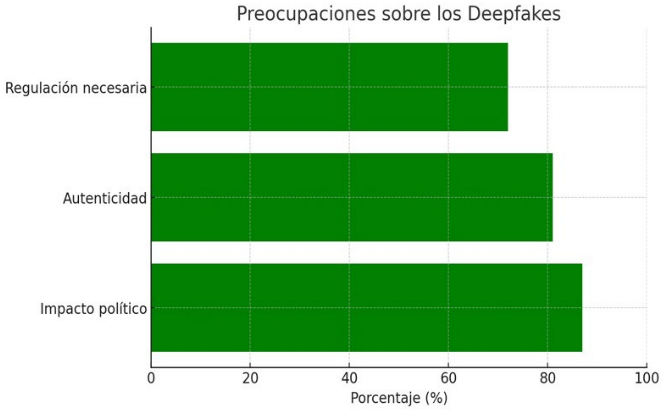


Source: Own elaboration, 2024

Truthfulness and Regulation: The following analyses are conducted: firstly, an examination of perceptions about the authenticity of content, and secondly, an analysis of opinions regarding the necessity for regulation of so-called “deepfakes”. Indeed, a unanimous consensus emerged among the participants of the focus group, with all respondents expressing an increased awareness of the potential risks associated with the veracity of information disseminated on Instagram. In the same instrument, political videos were perceived as manipulative by 92 per cent of the participants, thus highlighting a pattern of distrust towards this type of content. The remaining 8 per cent of the participants found the videos humorous.

With regard to the online questionnaire, Figure 2, it was reported by 81% of respondents that deepfakes give rise to concerns regarding the authenticity of the information. Moreover, a significant proportion of respondents, specifically 72%, advocate for stringent regulatory measures to be implemented.

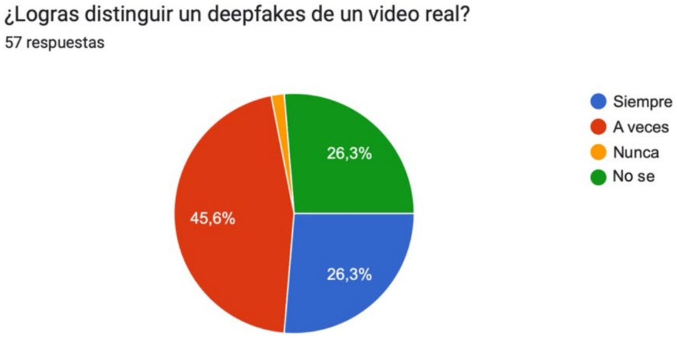
Figure 2. Alerts about Deepfakes



Source: Own elaboration, 2024

In a related study, less than half of participants (45.6%), when prompted with examples, were able to identify a deepfake from authentic video, as demonstrated in Figure 3.

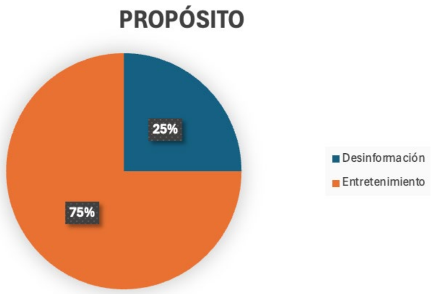
Figure 3. Deepfakes identification



Source: Own elaboration, 2024

Manipulation and purpose: An exploration of the digital alteration techniques employed in deepfakes and their impact on user perceptions. Nevertheless, the subject also encompasses the rationales underlying the conception of such video content. Of the eight selected fake videos with the most likes on Instagram Chile, 50% of them correspond to deepfakes with audio manipulation and 50% to videos with voice manipulation. Regarding the purpose of the videos, Figure 4 shows that 75% of videos are entertainment-focused, while only 25% are designed to disseminate disinformation.

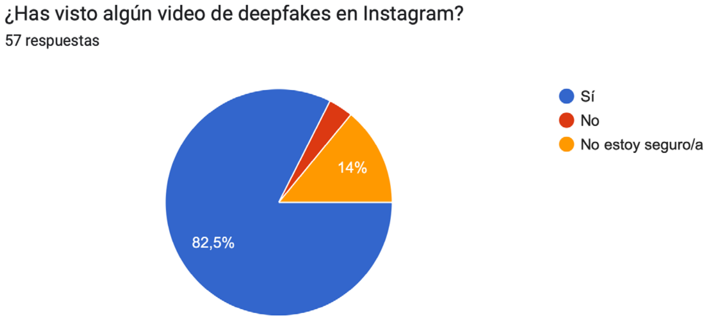
Figure 4. Purpose of deepfakes



Source: Own elaboration, 2024

User Interaction: The present category is concerned with the evaluation of user reaction and interaction with deepfakes on Instagram, with relevant metrics including likes, shares and comments. In the context of interaction, it is noteworthy that 82.5% of respondents have encountered a video of this nature on Instagram, as illustrated in Figure 5. A significant proportion of respondents, constituting 54.2%, perceive these images to be hazardous. However, a substantial proportion, amounting to 74.6%, have disseminated them on the social media platform Instagram.

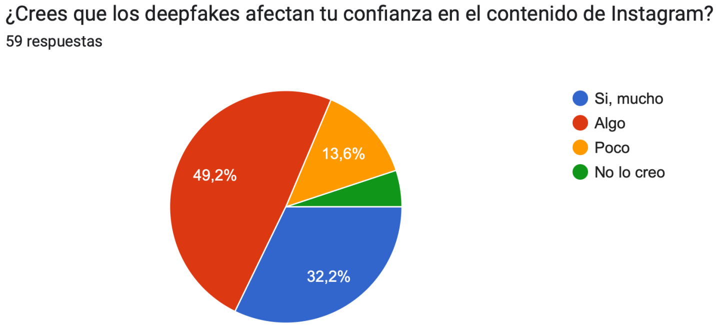
Figure 5. Interaction whit deepfakes



Source: Own elaboration, 2024

With regard to the impact of trust on the Instagram social network, 81.4% of users believe that it affects their trust in the social network to a great or moderate extent (see Figure 6). Within this same category, and with reference to the documentary analysis of Deepfakes, it was found that humorous videos engendered a greater number of positive interactions and likes. However, political videos attracted 89% of comments pertaining to concerns regarding misinformation.

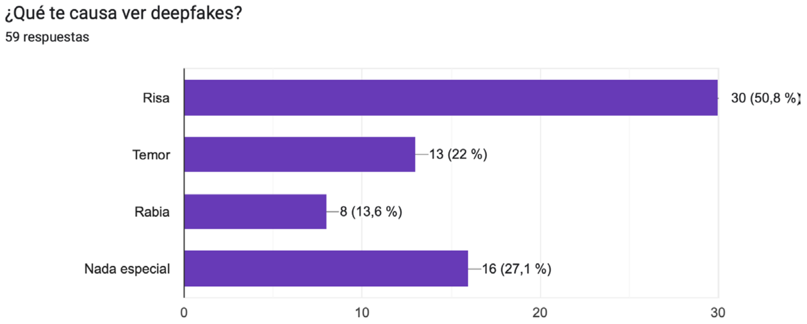
Figure 6. Trust in Instagram



Source: Own elaboration, 2024

Within the same category, laughter is the most common reaction among those who view this type of video on Instagram, with 50.8% of respondents expressing a preference for this response (see Figure 7). These data coincide with other instruments, such as the focus group, where participants mention feeling distrustful and detailing a fear of malicious use in political discourse. The purpose of this study is to examine the emergence of a new pattern in the field of deepfakes. The analysis will explore the general acceptance of deepfakes created for humorous purposes, and the blurred line between content that misinforms or manipulates perceptions.

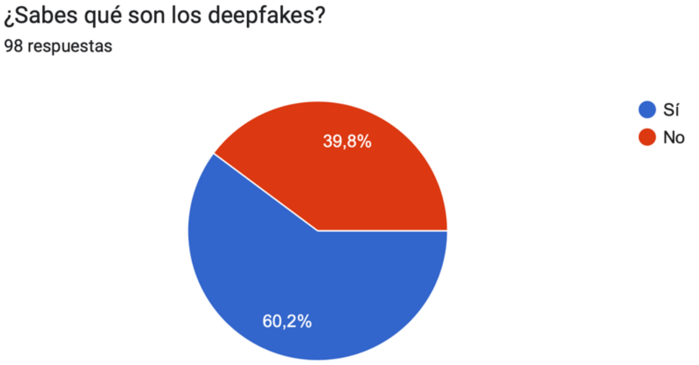
Figure 7. Perception before visualization



Source: Own elaboration, 2024

Knowledge of Deepfakes: This study examines the level of understanding users have about this technology and its implications, highlighting the relationship between prior knowledge and perceived trust. Regarding this category, Figure 8 shows that 60.2% of respondents are familiar with the concept of deepfakes.

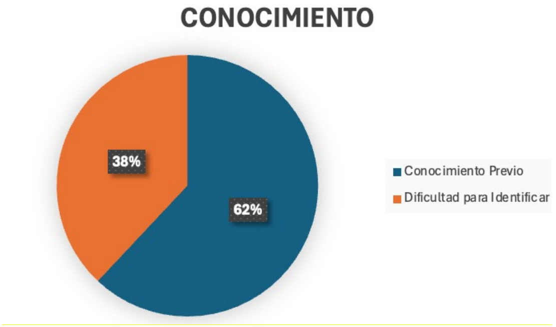
Figure 8. Conceptualization of deepfakes



Source: Own elaboration, 2024

With respect to the respondents' prior knowledge of this category of fake videos, Figure 9 shows that 62% of respondents reported awareness of the existence and circulation of such productions created by artificial intelligence, while 38% of them experienced difficulty in identifying them from authentic content. The results of the survey are very similar to Figure 8 regarding whether respondents have knowledge of deepfakes.

Figure 9 Prior knowledge



Source: Own elaboration, 2024

5. Conclusions

The advent of deepfakes, or fake videos, represents a significant emerging challenge in the digital sphere, given their capacity to manipulate the perception of reality. The increasing sophistication of these technologies, in combination with their ease of creation and difficulty of detection, necessitates rigorous preparation to mitigate their potential impact. It is a cause for concern that 80% of Instagram users have encountered Deepfakes videos, and that more than half of them, despite being aware of the fact that these videos have been created through artificial intelligence, have disseminated them without verifying their authenticity. It is important to note that such videos are not self-constructed; there is always a human behind the scenes directing the tools for a specific purpose. Whilst the majority of such content is created with the intention of entertaining the viewer and eliciting a positive response, such as laughter, there is a significant problem with the dissemination of misinformation through social networks, often disguised as humour. This is especially prevalent in videos featuring political leaders. This is a cause for concern when one considers that only 26% of users are able to distinguish a deepfake from a real video, thus rendering the possibility of the video being "believed" significant. A thorough analysis of deepfakes on Instagram discloses both the capabilities and the perils associated with this emergent technology.

While deepfakes have the potential to be a powerful creative tool in the field of entertainment and political satire, there is a significant risk that they could be used to spread disinformation. In particular, the phenomenon of “political deepfakes” poses a direct threat to the credibility of public figures and the trust placed in institutions. The capacity to manipulate voice and image has the potential to distort reality to such an extent that even the most discerning viewers may be susceptible to deception. Despite a considerable proportion of respondents being cognisant of the existence of this technology, the majority lack the tools or knowledge necessary to identify and counteract its impact. This underscores the pressing necessity to implement educational strategies that not only cultivate critical thinking but also foster digital literacy. It is imperative that such strategies are designed to empower users to detect manipulated content and to assess the authenticity of information sources.

In addressing the overarching objective of this research, the findings suggest that deepfakes generated by artificial intelligence exert a detrimental influence on the perception of users in the city of Concepción, Chile, with regard to the veracity of information disseminated on Instagram. This impact is reflected in a significant decline in user trust towards the platform, particularly in contexts where the accuracy and authenticity of information is paramount.

The results obtained through the tools employed during the fieldwork demonstrate the considerable impact of deepfakes on the trust of Instagram users. It was identified that most participants encounter difficulties in detecting manipulated content, which renders them vulnerable to misinformation, particularly in the political sphere and on issues related to public figures. This phenomenon is exacerbated by the inherent virality of social media, where content is rapidly disseminated without undergoing rigorous verification. Such users are not only less trusting but also call for clear regulations and strict sanctions to limit the malicious use of deepfakes. In this regard, digital platforms must assume a pivotal role in mitigating the deleterious impact of deepfakes. The data also revealed that users with a higher level of digital literacy exhibited a greater critical ability to identify deepfakes, thereby underlining the importance of implementing educational strategies that promote analysis and verification skills in the digital environment.

The implementation of advanced systems for detecting manipulated content, in conjunction with transparency initiatives that inform users about the provenance and authenticity of content, has the potential to play a pivotal role in restoring trust in these platforms. However, it is also crucial to emphasise the importance of users' own awareness of the content they are disseminating on social networks.

This research emphasises the shared responsibility of users, technology platforms and policymakers to address the ethical and social challenges posed by this emerging technology. It is imperative to recognise that the mitigation of the deleterious impact of deepfakes will only be attainable through concerted and collaborative endeavours. The establishment of a more secure, reliable and transparent digital environment is contingent upon such concerted action. Moreover, this study paves the way for future research that will explore specific strategies to combat the adverse effects of deepfakes in diverse cultural and social contexts, thereby consolidating a comprehensive approach to this phenomenon.

References

Ajder, H., Patrini, G., Cavalli, F., Cullen, L., & Deeptrace. (2019). The State of Deepfakes: landscape, threats, and impact.
In Deeptrace [Report]. https://regmedia.co.uk/2019/10/08/deepfake_report.pdf

Ballesteros, L., y Ruiz del Olmo, F. (2024). Vídeos falsos y desinformación ante la IA: El deepfake como vehículo de la posverdad. Revista De Ciencias De La Comunicación E Información, 29, 1–14. https://doi.org/10.35742/rcci.2024.29.e294

Citron, D. (2018, December 11). Deepfakes and the new disinformation war. Stanford CIS. https://cyberlaw.stanford.edu/publications/deepfakes-and-new-disinformation-war/

Citron, D. K., & Chesney, R. (n.d.). Deepfakes and the new disinformation war. Scholarly Commons at Boston University School of Law. https://scholarship.law.bu.edu/shorter_works/76/

Creswell, J. W. (1995). QUALITATIVE INQUIRY AND RESEARCH DESIGN. In Investigación Cualitativa Y Diseño Investigativo. https://academia.utp.edu.co/seminario-investigacion-II/files/2017/08/INVESTIGACION-CUALITATIVACreswell.pdf

Fletcher, R., & Park, S. (2017). The impact of trust in the news media on online news consumption and participation. Digital Journalism, 5(10), 1281–1299. https://doi.org/10.1080/21670811.2017.1279979

García-Ull, F. J. (2021). Deepfakes: el próximo reto en la detección de notícias falsas. Análisis, 64, 103. https://doi.org/10.5565/rev/analisi.3378

Hu K, Wang S, Lee D, Liu H (2020c) Mining disinformation and fake news: concepts, methods, and recent advancements. In: Disinformation, misinformation, and fake news in social media. Springer, Berlin, pp 1–19 https://doi.org/10.1007/978-3-030-42699-6_1

Hueso, L. (2022). Quién, cómo y qué regular (o no regular) frente a la desinformación. Teoría y realidad constitucional, (49), 199-238. https://dialnet.unirioja.es/servlet/articulo?codigo=8450040

Ibrahim, S. (2024, February 10). Cómo los 'deepfakes' están cambiando nuestra visión de la realidad. https://www.swissinfo.ch/spa/ciencia/cómo-los-ultrafalsos-cambian-nuestra-visión-de-la-realidad/46866008

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. Science, 359(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Martínez, E. (2023). Desinformación y fake news en TikTok: técnicas para su detección y prevención. https://oa.upm.es/76481/

Molina, O. (2024). Dissonance between technological transformation and journalistic ethics: A critical discussion of the impact of artificial intelligence on the media. adComunica Revista Científica De Estrategias Tendencias E Innovación En Comunicación, 91–114. 4 https://doi.org/10.6035/adcomunica.8002

Sierra Bravo, R. (1992). Técnicas de investigación social: teoría y ejercicios (p. 33)

Vaccari, C., & Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. Social Media + Society, 6(1). https://doi.org/10.1177/2056305120903408

Wardle, C., y Derakhshan, H. (2017). Desorden de la información: Hacia un marco interdisciplinario para la investigación y la formulación de políticas. Consejo de Europa. https://rm.coe.int/information-disorder-toward-an-interdisciplinary- framework-for-researc/168076277c

Wagner, A. [. (2024, September 13). Desinformación en la era digital. DIGITAL.CSIC. http://hdl.handle.net/10261/367658

Zuboff, S. (2019). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier ofPower (PublicAffairs,Ed.).ilustrada
https://www.hbs.edu/faculty/Pages/item.aspx?num=56791