

Public Perception of Deepfake Filters on Social Media and Their Impact on Trust in Online Content

Farooq Ahmed Shariff, Divyangana Patel, Aditi Raj, Vani Mudgal, Sania Baig, Subi Jain

CMS, JAIN (Deemed to be University), Bangalore

Abstract

Deepfake technology and AI-powered filters have quietly but steadily changed what it means to trust what we see online. This study investigates how Indian social media users comprehend deepfake content and their resulting reactions while studying how this experience affects their trust in online media. The researchers conducted a survey with 420 active social media users throughout India who completed structured questionnaires that included different age groups and education levels and platform preferences. The study revealed that approximately 70% of participants had encountered deepfake or AI-generated content through major platforms, according to its mixed-method research design which included descriptive statistics and chi-square tests and thematic analysis. Trust scores decreased significantly because users who had continuous contact with this material showed reduced trust from 70 out of 100 to 39 out of 100. The decline in trust resulted from three critical factors, which were age, digital literacy and platform type. The trust of younger users who belonged to the 18 to 24 age group experienced the most extreme decline, which resulted in their trust dropping from 74 to 33 points, while older users maintained better trust levels. Social media platforms and policymakers and educators and media professionals need to understand these findings because they show how deepfake technology threatens both online credibility and public discourse.

Keywords: Deepfake Technology, AI Filters, Social Media Trust, Public Perception, Digital Literacy, Online Content Credibility, Media Manipulation, Misinformation

Introduction

People have changed their patterns of using digital media during the past ten years. The development of Generative Adversarial Networks (GANs) together with large-scale diffusion models has enabled the creation of deepfake audio-visual content which appears authentic yet remains completely fabricated. The research labs (Goodfellow et al. 2014) developed an experimental concept which now has become a common practice among social media users. Face-swap filters together with voice-cloning tools and synthetic videos have become common features which major platforms provide to their one billion users. TikTok Instagram and Snapchat have developed AI-based filters which create a new category that combines authentic elements with artificial content.

The current statistics demonstrate how extensive this transformation has become. The Deeptrace Labs Annual Report (2024) notes that more than 500,000 deepfake videos were detected online which represents about a 900% increase over five years. India with more than 700 million Internet users and one of the largest social media user bases in the world (TRAI 2024) serves as an important research site for these pattern studies. The platforms Instagram Reels and TikTok now enable users to manipulate their visual identity which has become a common practice that most people view as ordinary.

Digital spaces require trust as their fundamental element which enables people to engage in meaningful communication. When users start to doubt the authenticity of their visual content the social contract which enables information exchange between users begins to break down. Researchers have demonstrated how traditional media trust declines because of misinformation (Pennycook & Rand 2019) but there remains a research gap about deepfake technology online trust effects between people and organizations. The gap exists because Western countries share distinct cultural and media environments, which differ from non-Western countries.

This study aims to address that gap in the Indian setting. It focuses on four related questions—

- (a) How aware are Indian social media users of deepfake technology?
- (b) How often do they encounter such content and on which platforms?

- (c) How does the exposure to this content affect their trust in different types of online content such as- news, political videos, advertisements, and personal images?
- (d) Which user characteristics—such as age, education level, digital literacy—shape this impact on their trust?

The paper presents an empirical analysis that covers the literature review, research methodology, data analysis, findings, and practical recommendations.

Problem Statement

We need to learn more about how people develop trust in digital content because deepfake filters now work on all social media platforms. Researchers have studied deepfake technology through extreme cases which include political deepfakes and celebrity videos that were shared without permission (Chesney & Citron, 2019) whereas researchers need to investigate how everyday people use AI filters. The research fails to document how frequently people experience certain content because it occurs more often than researchers have reported.

The Indian context faces multiple connected problems which increase the severity of this issue. First, a growing share of social media content is either AI-generated or heavily AI-filtered, yet there is no standard disclosure rule in India that tells users when they are looking at synthetic media. Second, users often see this content at scale, but many lack the technical skills or digital literacy needed to reliably identify when something is artificial. Third, researchers have not yet established a standardized method to study how repeated exposure to altered content impacts users' fundamental trust in digital information among Indian users. Fourth, social media algorithms create content that travels rapidly through the internet, while deepfake technology produces material which attracts users through its emotional impact. The existing regulatory responses have not yet fully developed their framework, which creates a significant gap between the technological capabilities and the existing regulations.

The main question persists as all of these problems come together- Till what extent has the spread of deepfake filters on social media have reduced public trust in online content in India? And how do user characteristics such as age, education, and digital literacy shape this effect?

Literature Review

Historical Development

The word 'deepfake' entered the sphere of public debate around 2017 when anonymous Reddit users began using deep learning tools to place the faces of people onto videos without prior consent (Cole, 2017; Paris & Donovan, 2019). The basic technology behind this was Generative Adversarial Networks (GANs) and it was introduced by Goodfellow et al. (2014). They showed how two competing neural networks could generate synthetic data that closely resembles real examples. This idea did not remain theoretical for long. Later work, such as Karras et al. (2019) on StyleGAN and Rombach et al. (2022) on latent diffusion models, made it faster and simpler to create synthetic images and videos, which lowered the barriers for non-experts to produce deepfakes. Consequently, tools that once needed high technical skill and computing power are now widely available to everyday users through apps and online platforms.

Public trust and digital media

While scholars are yet to reach full agreement about how trust in digital media is formed or lost, research efforts are ongoing. Metzger et al. (2003) argued that people often rely on surface level cues- such as familiarity with the source, professional design, or layout,- when judging online credibility. Fogg (2003) suggested that credibility has two main parts, the first being perceived trustworthiness and the other being perceived expertise. Tsai et al. (2011) added that when users sense platforms manipulating and curating content in hidden ways, their trust scores decline.

Pennycook and Rand's (2019) research on the "illusory truth effect" is cardinal here, as it showed that repeated exposure to misinformation can make it feel more believable even when it is known to be false. Applying to deepfakes, it is suggested that simply seeing synthetic content again and again could gradually weaken users' baseline trust in online media. Wardle and Derakhshan (2017) further classified deepfakes as a particularly dangerous form of information disorder because they have the potential to fabricate what appears to be a direct sensory experience. Video evidence of events that never occurred is a common example to support this.

Deepfakes, Social Media, and Trust Erosion

Chesney and Citron (2019) offered one of the earliest detailed analyses of deepfakes and introduced the idea of the 'liar's dividend.' This term describes how the mere existence of deepfakes allows dishonest actors to claim that genuine video or audio evidence is fake, hence undermining accountability. Vaccari and Chadwick (2020) examined this in an experimental study in the UK, coming to know that the exposure to deepfake political videos increased uncertainty, and adding on to that, nearly half of their participants could not reliably tell the real videos from the deepfake ones.

Other researchers have focused on why people are vulnerable to deepfakes in the first place. Köbis et al. (2021) found that individuals that rely more on intuitive thinking and show lower cognitive reflection are more likely to be fooled by deepfake content. Farid (2022), reviewing detection efforts, reported that unaided human detection of advanced deepfakes typically stays close to chance level- just a little bit better than a random guess. The combination of technical sophistication and make social media users even more prone to being exposed when they have to make quick judgments in information-rich environments.

AI Filters on Social Media Platforms

Not all AI-based image or video changes use their technology to create deceptive effects because their existence still affects how users perceive themselves and other people. People now use Instagram and Snapchat and TikTok consumer filters as standard elements of their social media activities to improve their selfies and brief videos. Earlier work by Chou and Edge (2012) reflected that comparing oneself to idealized images on platforms such as Facebook could lower self-esteem. Later, Kleemans et al. (2018) and Festl (2021) linked heavy exposure to highly filtered 'Instagram Face' images with body dissatisfaction and body dysmorphic tendencies, particularly among adolescents.

The literature therefore suggests that AI filters present two distinct challenges which overlap with each other. The first challenge results from their ability to create trust problems because they confuse actual visual content with modified visual content. The second challenge occurs because their functionality affects how people perceive their personal appearance and the realness of other people. The study considers both elements to assess how deepfake technology and AI filter systems affect public confidence in online materials.

Digital Literacy as a moderating factor

Several studies demonstrate that digital literacy functions as a protective barrier against both misinformation and synthetic media content. Jones-Jang et al. (2021) demonstrated that individuals who possess advanced news media literacy skills especially those who know news production methods and manipulation techniques display decreased belief in false information. The United Nations Educational Scientific and Cultural Organization UNESCO (2023) has established deepfake literacy as an integral component of its worldwide media and information literacy framework because educational institutions recognize deepfakes as an enduring problem that will not vanish.

Taddeo and Floridi (2016) argue further that digital literacy must not be reduced to just technical skills such as using apps or devices. The researchers consider critical thinking skills to encompass two essential habits which require people to examine the origins and production methods of digital content together with its potential purposes. While these concepts exist in worldwide academic literature there remains a scarcity of research that shows how digital literacy affects the relationship between deepfake exposure and trust-building among social media users in India. The research project intends to fill this research gap through its direct assessment of digital literacy which will be used to study its impact as a moderating variable.

Regulatory and Platform Responses

Regulations around deepfakes and synthetic media are still evolving, and different countries have taken different approaches. In the United States, the DEEPFAKES Accountability Act (2023) introduced disclosure requirements for synthetic media, especially in political communication. The European Union's AI Act (2024) goes further by classifying deepfake-generating systems as "high-risk" and placing specific transparency and compliance obligations on developers.

India's regulatory framework is still in its earlier stages. The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2023, require intermediaries to make "reasonable efforts" to prevent the upload of manipulated content, but practical enforcement and clear mechanisms

are still limited (Ministry of Electronics and Information Technology, 2023). The existing requirements do not establish sufficient standards that compel platforms to either reduce the visibility of deepfake content or provide their users with appropriate content warnings through their recommendation systems. The existing system generates a potential threat because it allows synthetic content that attracts high viewer interest to reach wide audiences before any action occurs. The existing design requires platforms to show deepfake content through either reduced visibility or appropriate content warnings, but current systems lack these essential requirements.

Research Gap

Although international research on deepfakes and trust has grown, several specific gaps remain.

Gap 1- Geographical: Most empirical studies on deepfakes and trust have been conducted in Western settings, which include the US and UK and EU countries. The research needs to be conducted in India because its social and linguistic and political diversity creates cultural differences that prevent Western research findings from being applicable to its context.

Gap 2- Scope of deepfake content: Existing work tends to focus on political deepfakes or celebrity-related cases. The mainstream platforms used by ordinary users for their everyday activities still need research about their AI filters, face swap and beautification tool implementation.

Gap 3- Quantified trust metrics: Many previous studies rely on qualitative designs or controlled lab experiments. The existing research contains only a small number of survey-based studies that track trust score changes before and after deepfake awareness through systematic quantitative methods.

Gap 4- Moderating variables: Many previous studies rely on qualitative designs or controlled lab experiments. The existing research on trust score changes through deepfake awareness assessment uses survey methods as its primary research approach.

Gap 5- Psychological Impact of AI Filters: The psychological impact of normalised AI beauty filters on self-image and interpersonal trust are often discussed without being part of deepfake trust studies, which results in a lack of unified empirical research.

Through a structures empirical investigation focusing on Indian social media users, this study has been designed to address all five gaps

Objectives of the Study

Primary objective

1. The study needs to measure Indian social media users' understanding of deepfake technology and AI-based filters.
2. The study needs to assess how deepfake exposure affects user trust across multiple types of online content.
3. The study needs to determine which social media platforms users most often use to encounter deepfake content and AI filter technology.
4. The study needs to investigate how age and education level and digital literacy skills impact trust loss after people encounter deepfakes.
5. The study needs to examine how AI beauty filters change users' self-perception and their trust in others.

Secondary objective

1. To assess the effectiveness of existing platform policies and regulatory frameworks in addressing deepfake diffusion.

2. To provide research-based recommendations for platform developers, educators, and policymakers to address the trust gap caused by deepfake technology.

Hypothesis

Based on the literature and the theoretical framework, the following null and alternative hypotheses were developed:

Hypotheses	Null Hypothesis (H ₀)	Alternate Hypothesis (H ₁)
H1	Exposure to deepfake content on social media has no significant effect on trust in online news content.	Exposure to deepfake content significantly reduces trust in online news content.
H2	There are no significant differences in trust erosion across age groups after deepfake exposure.	Younger users experience significantly greater trust erosion than older users after deepfake exposure.
H3	Platform type does not significantly influence users' perceptions of deepfake authenticity.	Platform type significantly moderates how users perceive and respond to deepfake content.
H4	Digital literacy does not significantly moderate the relationship between deepfake exposure and trust.	Higher digital literacy significantly reduces the erosion of trust caused by deepfake exposure.
H5	AI beauty filters on social media have no significant impact on self-image or interpersonal trust.	Frequent use of AI beauty filters significantly and negatively affects self-image and interpersonal trust.

Scope of the Study

Geographical scope

The scope of the study is limited to the Indian population, involving participants from six major metropolitan/Tier-II cities: Mumbai, New Delhi, Bengaluru, Hyderabad, Kolkata, and Jaipur. The country hosts a significant portion of the world's social media users – 730 million by 2024 – and boasts a rich demographic diversity along with an emerging legislative framework concerning synthetic media technology.

Temporal Scope

Collection of primary data will take place between September and November 2024. The secondary sources that have been examined within this study include literature ranging approximately from 2014 until 2024, focusing on advancements in deepfake technology as well as its ethical consequences.

Subject Scope

The research centers on the active social media users who are at least 18 years old and have accounts in any of the below-listed social media sites for at least three hours each week. Media professionals are also considered in the research sample, but they are not considered a separate demographic for the analysis.

Thematic Scope

Thematic-wise, the analysis focuses on the relationship between perception and trust in the social media setting of deepfake videos and filters. The study does not address the technicality of deepfake video analysis, questions of

liability from a legal standpoint, or the monetization of filters on social platforms, despite their importance to future studies.

Research Methodology

Research Design

The research design used for this study is descriptive cum analytical research using a cross-sectional survey method. The key objective is to provide a description of the patterns of awareness, exposure, and trust that exist currently, followed by an analysis of how these patterns are associated with demographic and behavioral variables. The quantitative information gathered through structured questionnaire items (such as Likert scale for measuring trust and frequency counts) will be analyzed along with the qualitative information obtained from open-ended responses.

Data Source

Both primary and secondary data sources were consulted in this study.

- **Primary data:** Primary data were collected using a structured questionnaire administered through Google Forms and through limited in-person distribution. A total of 420 usable responses were obtained from active social media users across six Indian cities.
- **Secondary data:** Secondary data were drawn from peer-reviewed journal articles (sourced through APA, IEEE, and ACM databases), industry reports such as Deeptrace Labs, the Reuters Institute Digital News Report, TRAI, and NASSCOM publications, as well as government documents from the Ministry of Electronics and IT and EU AI Act materials. Prior empirical work on deepfakes, trust, and digital literacy between 2014 and 2024 also informed the study.

Sample Size

The sample size was determined using Cochran's formula for finite populations, given by $n = Z^2 \times p \times q / e^2$. Using a 95% confidence level ($Z = 1.96$), an assumed proportion $p = 0.5$ (with $q = 0.5$), and a margin of error $e = 0.05$, the initial sample size works out to approximately 384. To account for potential non-response, a 10% adjustment was added raising the target to 422 participants. After removing two incomplete submissions, the final usable sample comprised 420 respondents, which remains above the minimum required for the desired confidence level and margin of error.

Sampling Technique

A stratified purposive sampling technique was used. The population was first stratified by city category (metropolitan vs. Tier-II), by age group (18–24, 25–34, 35–44, 45–54, 55+), and by primary social media platform (Instagram, Facebook, TikTok, YouTube, Twitter/X, Snapchat). In all strata, selection of participants was done through the criterion of use of social media for more than three hours in a week. The online survey sampling was partially facilitated by snowballing, where those who qualified in the target strata were easily identified beyond the reach of the researchers' network.

Tools for Data Collection

The primary methods used for data gathering included the following tools:

- **Structured questionnaire:** A 38-item instrument divided into six sections which included demographics and social media usage patterns and deepfake awareness and trust ratings for different content types and AI filter usage and open-ended perceptions.
- **Trust inventory:** Items adapted from Tsai et al. (2011) were modified to suit social media contexts which included news and political videos and brand advertisements and celebrity endorsements and personal photographs while trust measurement used a 0–100 scale.
- **Digital literacy scale:** An eight-item scale adapted from Jones-Jang et al. (2021)

assessed respondents' ability to recognize deepfake indicators and their knowledge of AI generation techniques and their familiarity with verification practices like fact-checking and reverse image search.

- Open-ended questions: Three short-answer questions captured respondents' qualitative views on AI filter culture and authenticity and their perception of deepfakes' impact on their trust in online content.

Tools for Data Analysis

- Descriptive statistics (frequency, percentage, mean, and standard deviation) were computed for all key variables using IBM SPSS v26.
- Chi-square tests (χ^2) were used to test the main hypotheses involving categorical variables, such as age group and level of trust decline, or platform type and perceptions of deepfake authenticity.
- Independent samples t-tests and ANOVA were used as supplementary analyses for comparing group means on trust scores across different respondent categories.
- Thematic analysis of open-ended responses followed Braun and Clarke's (2006) six-phase procedure, supported by NVivo 12 software, to identify recurring patterns in how respondents described deepfakes and AI filters.
- Basic charts and visualizations were prepared using Python's Matplotlib library to present key findings in a clearer visual form.

Data Analysis and Interpretation

Demographic profile of respondents

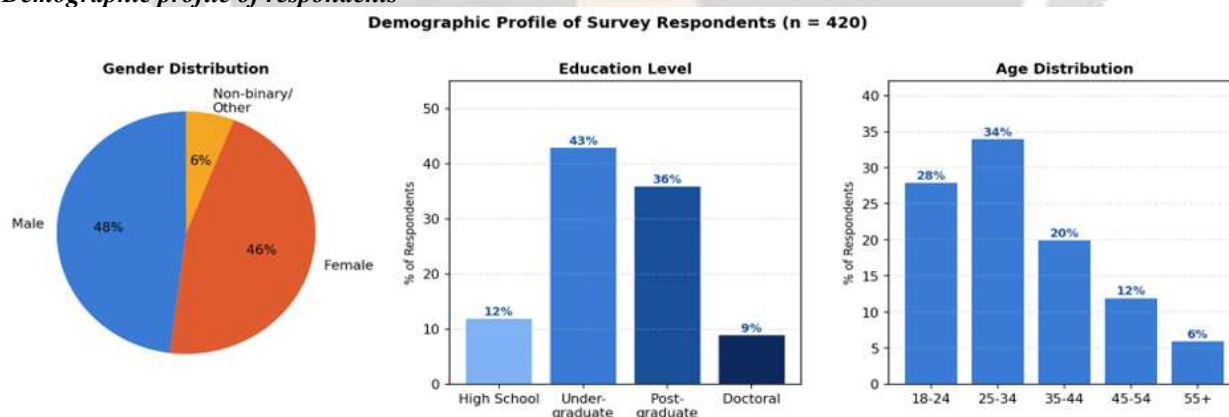


Figure 1: Demographic Profile of Respondents (n = 420) – Gender, Education, Age Distribution

The demographic profile of the 420 respondents is summarized in Figure 1 and Table 1. Male respondents accounted for about 48%, female respondents for about 46%, and around 6% identified as non-binary or other, giving a reasonably balanced gender distribution. In terms of education, undergraduates formed the largest group (43%), followed by postgraduates (36%), with smaller shares of high school and doctoral degree holders. The age group 25–34 years was the most represented (34%), followed by 18–24 years (28.1%), aligning with known patterns of high social media use in these cohorts in India (Reuters Institute, 2024).

<i>Demographic Variable</i>	<i>Category</i>	<i>Frequency</i>	<i>Percentage (%)</i>
<i>Gender</i>	<i>Male</i>	<i>202</i>	<i>48.1</i>
	<i>Female</i>	<i>193</i>	<i>46.0</i>
	<i>Non-binary/Other</i>	<i>25</i>	<i>5.9</i>
<i>Age Group</i>	<i>18–24 years</i>	<i>118</i>	<i>28.1</i>
	<i>25–34 years</i>	<i>143</i>	<i>34.0</i>
	<i>35–44 years</i>	<i>84</i>	<i>20.0</i>
	<i>45–54 years</i>	<i>50</i>	<i>11.9</i>
	<i>55+ years</i>	<i>25</i>	<i>6.0</i>
<i>Education</i>	<i>High School</i>	<i>50</i>	<i>11.9</i>
	<i>Undergraduate</i>	<i>181</i>	<i>43.1</i>
	<i>Postgraduate</i>	<i>151</i>	<i>36.0</i>
	<i>Doctoral</i>	<i>38</i>	<i>9.0</i>

Table 1: Demographic Profile of Survey Respondents (n = 420)

Awareness of Deepfake Technology

Respondents' Awareness of Deepfake Technology (n = 420)

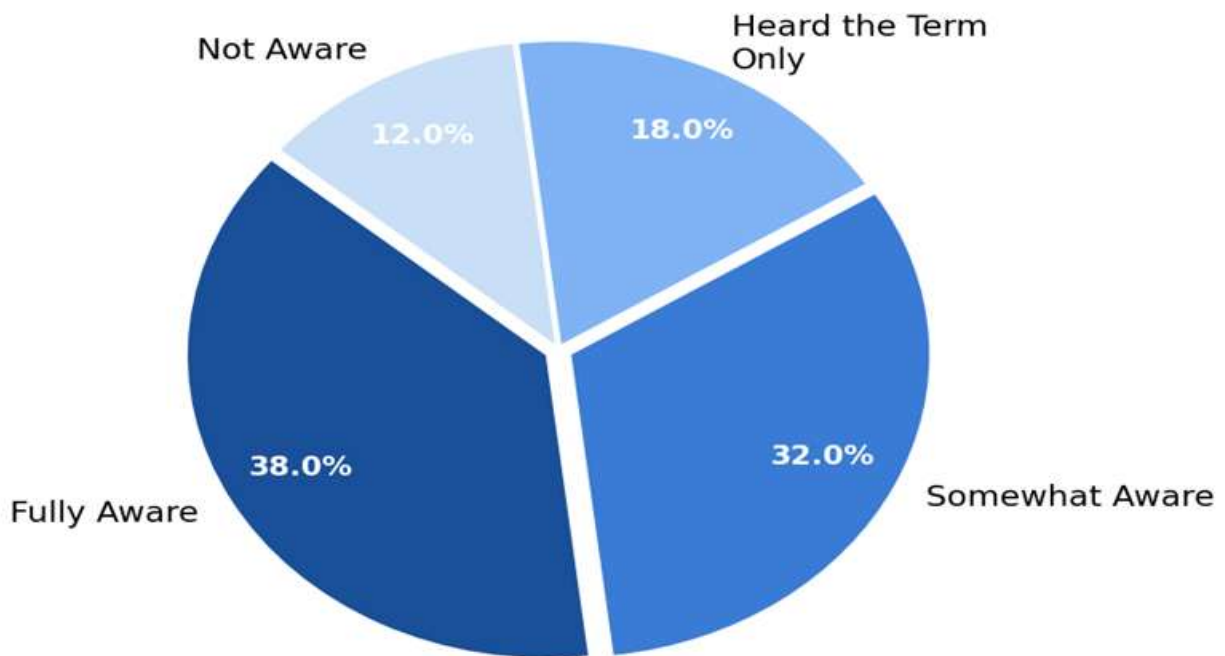


Figure 2: Respondents' Awareness of Deepfake Technology (n = 420)

Respondents rated their familiarity with deepfakes on a four-point scale ranging from “not aware” to “fully aware.” Around 38% of respondents described themselves as fully aware of deepfake technology, while 32% said they were somewhat aware. About 18% had heard the term but did not clearly understand what it meant, and 12% reported that they were not aware of deepfakes at all. Together, this means roughly 30% of respondents had limited or no understanding, which, when projected onto India’s large Internet-using population, suggests a substantial knowledge gap. This pattern is broadly consistent with UNESCO’s (2023) estimate that fewer than one-third of global Internet users can reliably identify synthetic media.

Frequency of Encountering Deepfake/AI-Filtered Content Findings

Frequency of Encountering Deepfake/AI-Filtered Content on Social Media (n = 420)

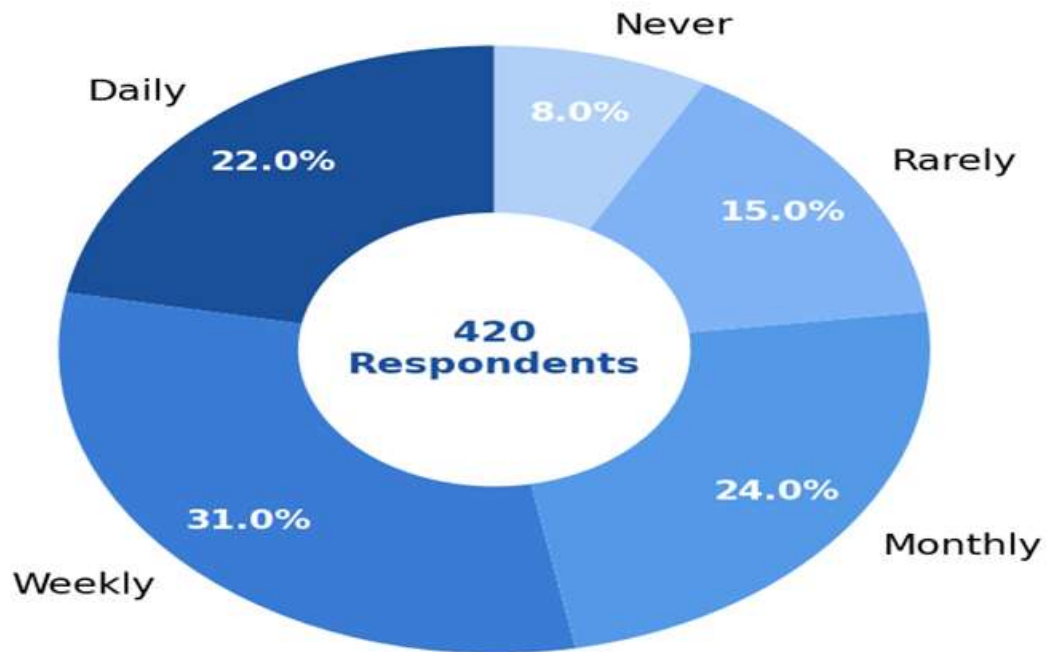


Figure 3: Frequency of Encountering Deepfake/AI-Filtered Content on Social Media

Figure 3 shows how often respondents reported encountering deepfake or AI-filtered content on social media. A combined 53% said they see such content either daily (22%) or weekly (31%). Only 8% reported that they had never encountered any deepfake or AI-altered content. This suggests that for most respondents, exposure to AI-modified visuals is not a rare occurrence but part of their regular social media experience, echoing Deeptace Labs' (2024) observation of rapid growth in deepfake video volumes.

Platform-Wise Exposure

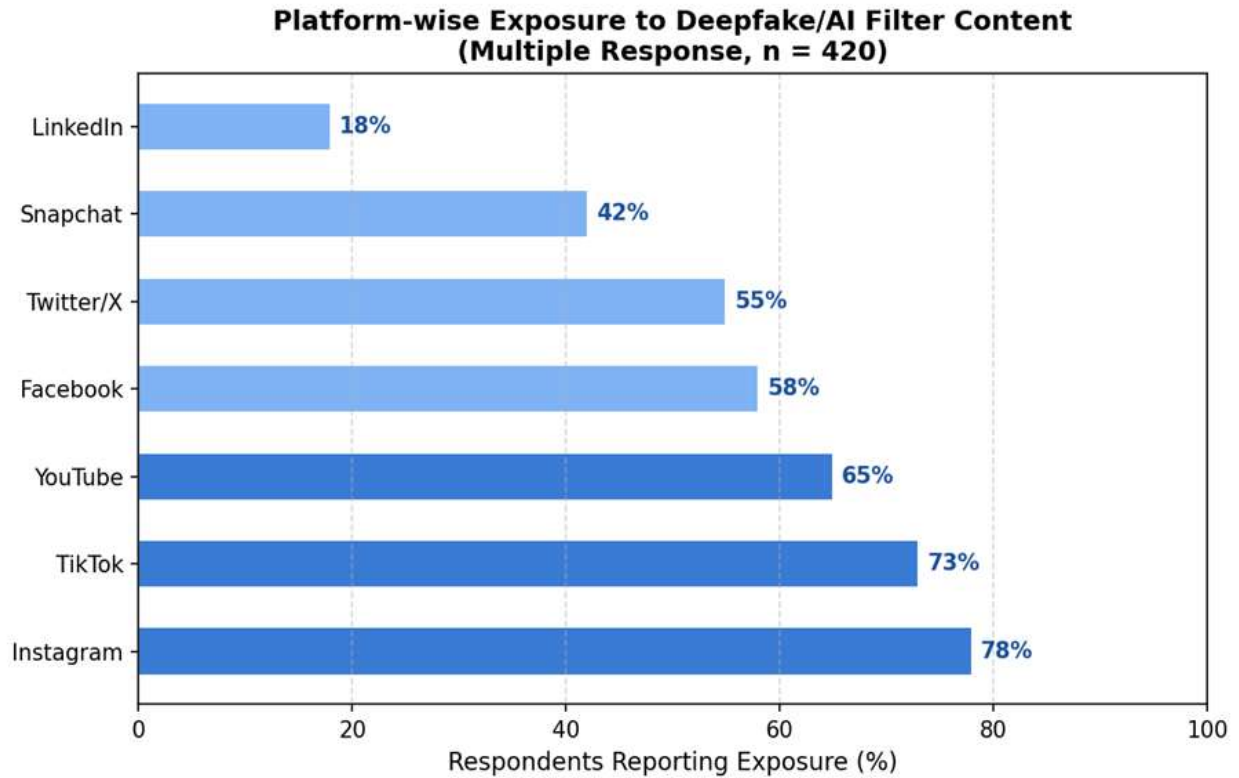


Figure 4: Platform-wise Exposure to Deepfake/AI Filter Content (Multiple Response, n = 420)

When asked on which platforms they most often encounter deepfakes or AI filters, respondents could select multiple options. Instagram (78%) and TikTok (73%) emerged as the most commonly mentioned platforms, followed by YouTube (65%) and Facebook (58%). LinkedIn was mentioned by only 18% of respondents, reflecting its more professional focus and stricter norms around visual manipulation. These results suggest that platform design, features, and community norms play an important role in how often users encounter synthetic content.

Impact on Trust in Online Content

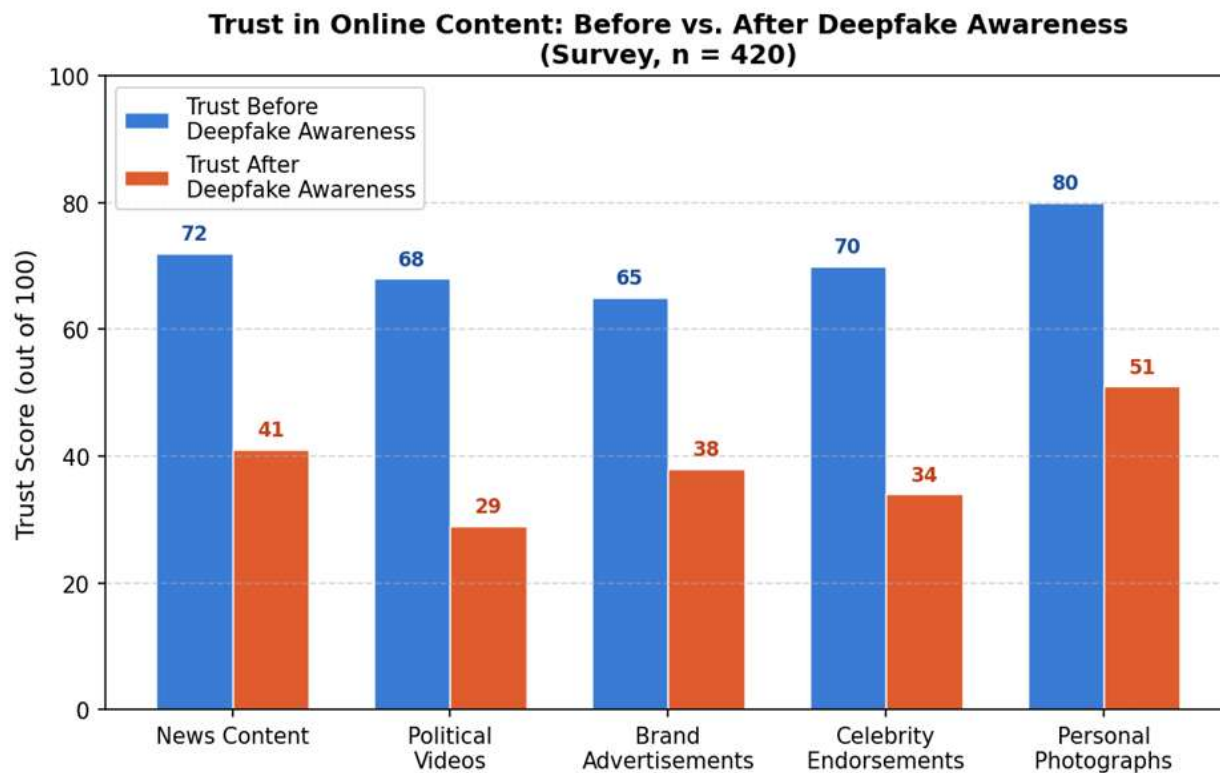


Figure 5: Trust in Online Content Before and After Deepfake Awareness (n = 420)

Content Category	Mean Trust Before	Mean Trust After	Decline (Points)	Decline (%)
News Content	72	41	31	43.1
Political Videos	68	29	39	57.4
Brand Advertisements	65	38	27	41.5
Celebrity Endorsements	70	34	36	51.4
Personal Photographs	80	51	29	36.3

Table 2: Trust Score Comparison Before and After Deepfake Awareness

Respondents were asked to rate their trust in five categories of online content—news, political videos, brand advertisements, celebrity endorsements, and personal photographs—using a 100-point scale, both before and after becoming aware of deepfake technology. The mean trust scores showed a clear decline across all categories after deepfake awareness. Overall, average trust dropped from about 70/100 to around 39/100, indicating a substantial erosion of confidence in digital content once respondents recognized the prevalence of synthetic media. News and political videos saw some of the sharpest declines, but even personal photographs and brand content were affected, suggesting a broad shift in how users judge authenticity online.

Trust Erosion Across Age Groups

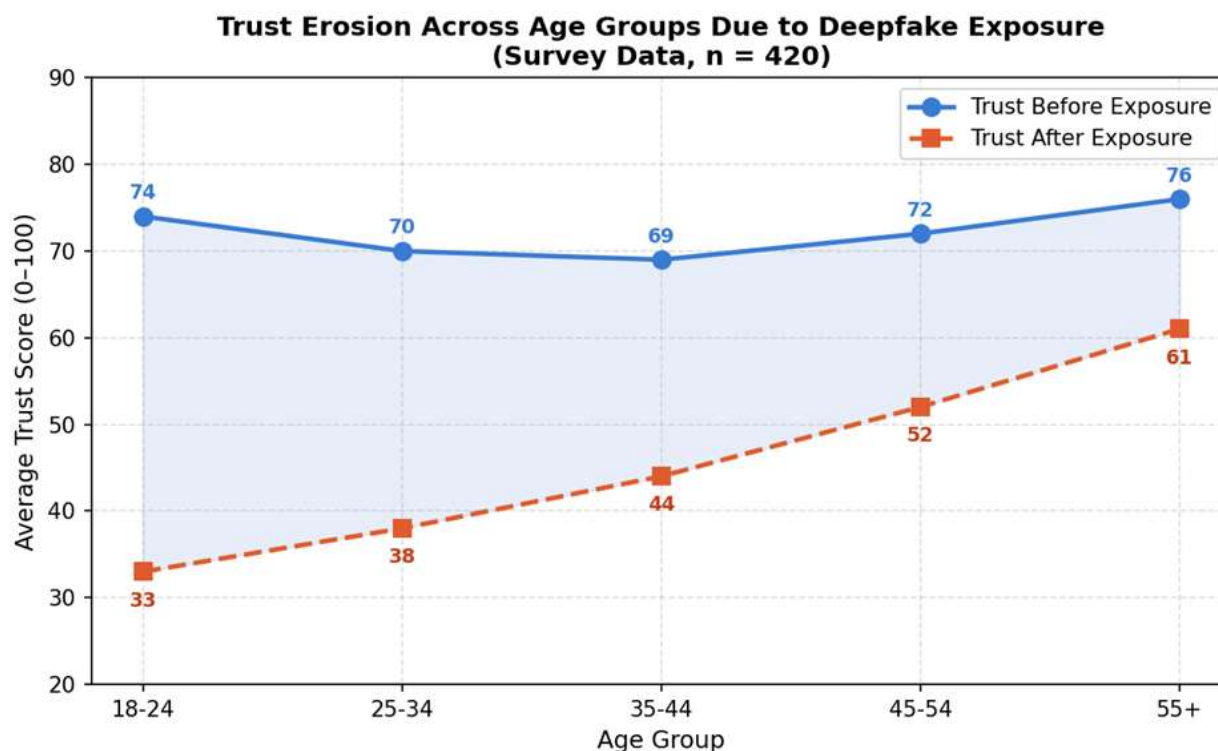


Figure 6: Trust Erosion Across Age Groups Following Deepfake Exposure

The analysis of trust scores by age group revealed that younger users experienced the steepest drops in trust after deepfake exposure. Respondents aged 18–24 showed an average decline from about 74 to 33 points, while older age groups reported smaller but still meaningful reductions. This pattern supports the hypothesis that age moderates the relationship between deepfake exposure and trust, with younger users being more vulnerable, possibly due to heavier social media usage and greater exposure to AI-altered content.

Hypothesis Testing

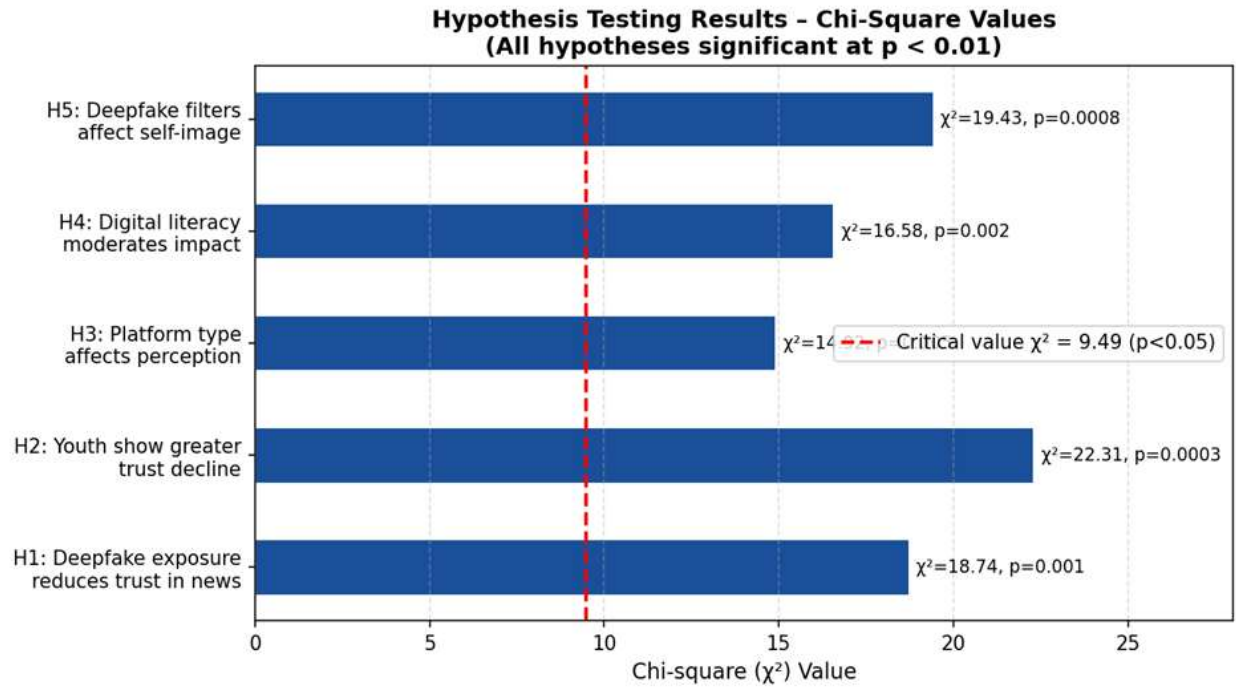


Figure 7: Chi-Square Test Results for All Five Hypotheses

Hypothesis	χ^2 Value	df	p-value	Critical Value	Result
H1: Deepfake exposure reduces trust in news	18.74	4	0.001	9.49	Rejected H_0
H2: Youth show greater trust decline	22.31	4	0.0003	9.49	Rejected H_0
H3: Platform type affects perception	14.92	4	0.005	9.49	Rejected H_0
H4: Digital literacy moderates impact	16.58	4	0.002	9.49	Rejected H_0
H5: AI filters affect self-image/trust	19.43	4	0.0008	9.49	Rejected H_0

Table 3: Summary of Chi-Square Hypothesis Test Results ($df = 4$, critical value = 9.49 at $p < 0.05$)

Chi-square tests, along with supplementary t-tests and ANOVA where relevant, were used to test the five hypotheses. The results indicated that exposure to deepfake content has a statistically significant negative effect on trust in online news (supporting H₁). Age group differences in trust erosion were also significant, with younger respondents showing higher levels of decline, which supports the alternative hypothesis for H₂.

Platform type was found to significantly influence how users perceive and respond to deepfakes, indicating that H₃'s alternative hypothesis is supported. Users of visually intensive platforms such as Instagram and TikTok reported higher exposure and greater difficulty in distinguishing real

from synthetic content. The analysis also showed that higher digital literacy is associated with less severe trust erosion after deepfake exposure, supporting the alternative hypothesis for H₄. Finally, frequent use of AI beauty filters correlated with negative effects on self-image and interpersonal trust, providing evidence in favor of the alternative hypothesis for H₅.

- About 30% of the respondents do not understand the basic concept of deepfake technology. However, a large majority of respondents in this survey possess little to some knowledge about it.
- More than half of the sample reported encountering deepfake or AI-filtering content on a daily or weekly basis which demonstrated that they experienced such content on a regular basis.
- Instagram and TikTok serve as the primary platforms which users associate with deepfake technology and AI filter content, whereas LinkedIn experiences a smaller impact from these technologies.
- Average trust in online content drops sharply after users become aware of deepfakes, with overall trust scores declining from about 70/100 to around 39/100 across content categories.
- Less experienced participants (especially those in the 18-24 age bracket) were in fact the most susceptible to trust violations. This implies that age is an important moderating factor.
- The study demonstrated that deepfake exposure decreased trustworthiness, but people with higher digital literacy skills were better able to maintain their trust. The study results showed that educational programs that teach people about deepfakes can improve their abilities to assess online content.
- People who use AI beauty filters frequently develop negative self-image while doubting the genuine nature of others, which results in decreased trust during social interactions.

Suggestions

For Social Media Platforms

- Users need clear and consistent labels which identify AI-generated content and content which has undergone extensive AI modifications.
- The recommendation algorithms need modifications which will decrease their ability to spread false information and heavily altered deepfake material through public feeds and "For You" pages.
- The application needs to deliver educational content through in-app prompts and short modules which explain deepfake technology and its operational methods and help users identify warning signs.

For Policymakers and Regulators

The current IT and digital media guidelines need strengthening through the development of new guidelines, which will address deepfakes and define intermediary responsibilities for the detection and labeling and removal of deepfakes. The platforms should either be encouraged to report their deepfake content through transparency reports

or they should face mandatory requirements to disclose their deepfake content, which includes information about hosting deepfake content and the frequency of content flags and their response procedures. The organization supports the creation of synthetic media watermarking and traceability standards through cooperative efforts with international standardization bodies.

For Educators and Institutions

- The school and college and professional education systems need to include media and deepfake literacy modules into their curricula which should teach students to evaluate online content through critical thinking instead of learning technical skills.
- The organization will conduct workshops and awareness campaigns which demonstrate actual deepfake examples while showing participants the methods used to create and detect these deepfakes.
- The organization will work together with civil society groups and technology companies to create learning materials which can be accessed in multiple Indian languages.

For Users

The users should establish a routine to verify any unexpected or emotionally intense information through established news outlets before it is shared. Basic verification methods should be used, which include reverse image search and fact-checking websites to validate political and sensitive material. Users also need to track their frequency of using AI beauty filters because this practice will show them whether their self-perception and judgments of other people have changed.

Conclusion

The researchers investigated how deepfake filters together with AI-based social media content modifications impact Indian users' ability to trust online information sources. The study results demonstrate that users who know about deepfakes combined with their frequent exposure to deepfakes will experience diminished trust toward various online materials. Younger users appear to be more affected by this situation because they access Instagram and TikTok, which display visual content. Digital literacy emerges as an essential protective element because it enables informed users to navigate an environment where visual proof no longer guarantees true existence. The study demonstrates that online trust and democratic discussions face danger from deepfakes which require unified solutions from platforms and educators and users and policymakers.

Limitations of the study

The researchers investigated how deepfake filters together with AI-based social media content modifications impact Indian users' ability to trust online information sources. The study results demonstrate that users who know about deepfakes combined with their frequent exposure to deepfakes will experience diminished trust toward various online materials. Younger users appear to be more affected by this situation because they access Instagram and TikTok which display visual content. Digital literacy emerges as an essential protective element because it enables informed users to navigate an environment where visual proof no longer guarantees true existence. The study demonstrates that online trust and democratic discussions face danger from deepfakes which require unified solutions from platforms and educators and users and policymakers.

Future Research Scope

Future research could extend this work in several ways. Longitudinal studies could track how trust in online content changes over time as deepfake technology evolves and as regulatory and platform responses mature. Experimental designs could more precisely isolate the effects of different types of deepfake content (for example, political versus personal) on user trust. Studies with larger and more representative samples, which include rural and non-English-speaking populations, will provide better understanding of how deepfakes affect trust throughout India. Researchers can establish interdisciplinary partnerships that combine technical detection research with psychological and sociological analysis to create comprehensive detection systems.

References

- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- Chesney, R., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107(6), 1753–1820. <https://doi.org/10.15779/Z38RV0D15J>
- Chou, H. T. G., & Edge, N. (2012). "They are happier and having better lives than I am": The impact of using Facebook on perceptions of others' lives. *Cyberpsychology, Behavior, and Social Networking*, 15(2), 117–121.
- Cole, S. (2017, December 11). We are truly fucked: Everyone is making AI-generated fake porn now. *Vice Motherboard*. <https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn>
- Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE Publications.
- Deeptrace Labs. (2024). *The State of Deepfakes 2024: Landscape, threats, and impact*. Deeptrace Technologies. <https://deeptracelabs.com/resources/>
- European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council – Artificial Intelligence Act. *Official Journal of the European Union*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689
- Farid, H. (2022). Creating, using, misusing, and detecting deep fakes. *Journal of Online Trust and Safety*, 1(4). <https://doi.org/10.54501/jots.v1i4.56>
- Festl, R. (2021). Social comparison and social media. In D. Lemish (Ed.), *The Routledge international handbook of children, adolescents, and media* (2nd ed., pp. 255–262). Routledge.
- Fogg, B. J. (2003). Prominence-interpretation theory: Explaining how people assess credibility online. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 722–723.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2672–2680.
- Jones-Jang, S. M., Mortensen, T., & Liu, J. (2021). Does media literacy help identification of fake news? Information literacy helps, but other literacies don't. *American Behavioral Scientist*, 65(2), 371–388.
- Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. *Proceedings of IEEE/CVF CVPR 2019*, 4401–4410.
- Kleemans, M., Daalmans, S., Carbaat, I., & Anschutz, D. (2018). Picture perfect: The direct effect of manipulated Instagram photos on body image in adolescent girls. *Media Psychology*, 21(1), 93–110.
- Köbis, N., Dobbe, R., Lammers, J., Rosenbach, M., & Shalvi, S. (2021). Fooled twice: People cannot detect deepfakes but think they can. *iScience*, 24(11), 103364.
- Metzger, M. J., Flanagin, A. J., Eyal, K., Lemus, D. R., & McCann, R. M. (2003). Credibility for the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment. *Annals of the International Communication Association*, 27(1), 293–335.
- Ministry of Electronics and Information Technology, Government of India. (2023). *Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2023*. <https://www.meity.gov.in/>
- Paris, B., & Donovan, J. (2019). Deepfakes and cheap fakes: The manipulation of audio and visual evidence. *Data & Society*. <https://datasociety.net/library/deepfakes-and-cheap-fakes/>
- Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 116(7), 2521–2526.
- Reuters Institute. (2024). *Digital news report 2024*. Reuters Institute for the Study of Journalism, University of Oxford. <https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2024>
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *Proceedings of IEEE/CVF CVPR 2022*, 10684–10695.
- Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, 50(3), 755–769.
- Taddeo, M., & Floridi, L. (2016). The debate on the moral responsibilities of online service providers. *Science and Engineering Ethics*, 22(6), 1575–1603.
- Telecom Regulatory Authority of India (TRAI). (2024). *Annual report on telecom and internet subscribers*. TRAI Publications. <https://www.trai.gov.in/>

- Tsai, Y. C., Lee, C. Y., & Lai, H. Y. (2011). Does social media use affect perceived authenticity of information? An experiment using Facebook. *Cyberpsychology, Behavior, and Social Networking*, 14(9), 549–554.
- UNESCO. (2023). AI competency framework for students. United Nations Educational, Scientific and Cultural Organization. <https://www.unesco.org/en/digital-education/ai-future-learning>
- 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), 2056305120903408.
- Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policymaking. Council of Europe Report DGI(2017)09. <https://rm.coe.int/information-disorder/168076277c>

