Contents lists available at International Journal Press

Journal of Digital Justice and Legal Innovation

journal homepage: https://ijpress.com/jdjli

Volume 1 · Issue 1 · 2026 pp. 11-18

Corporate Vulnerability in the Era of Synthetic Media: Challenges of AI-Generated Manipulation

Saumya Dwivedi[1,*] Mohit Bhatnagar[2]

[1] *Research Scholar, Amity Law School, Amity University, Noida, Uttar Pradesh, India*

[2] *Advocate, Karkardooma District Court, Delhi, India*

* Corresponding author: dwivedisaumya1318@gmail.com

ARTICLE INFO

Article history:

Received: 23-07-2026

Received in revised form: 03-08-2026

Accepted: 05-08-2026

Available online: 05-08-2026

Keywords:

Corporate Misinformation

Deepfake Manipulation

Generative AI Risk

India Digital Governance

Synthetic Media

ABSTRACT

Corporate trust has long rested on the assumption that what is seen and heard can be believed. Generative AI, which now produces photorealistic video, cloned audio and fabricated text at low cost and high speed, has unsettled that assumption. This paper examines the vulnerabilities that synthetic media creates for corporate organisations: investor fraud, brand defamation, executive impersonation and coordinated disinformation. Drawing on documented incidents in the United States, Europe and India, on crisis communication and information security scholarship, and on a comparative reading of regulatory initiatives across jurisdictions, it argues that most organisations are poorly equipped for attacks that bypass the conventional cyber security perimeter. Three themes are developed: the anatomy of synthetic media attacks on corporate targets; the internal governance failures that compound institutional vulnerability; and the practical limits of current detection technology and law. Particular attention is given to the Indian context, whose combination of deep digital penetration, constrained regulatory enforcement and a fast-growing corporate sector produces a distinctive and under-researched risk profile. Recommendations are directed at corporate leadership, communications practitioners and policymakers. The analysis concludes that synthetic media manipulation is not merely a communications problem but a structural governance risk that will deepen as generative capability advances.

1. INTRODUCTION

Reputational attacks have been a long-standing practice in corporate institutions, and with the advent of large-scale generative AI, the cost and ease of deception have become more accessible and more costly than its defense. What used to take weeks to prepare with production resources, trained actors and others can be done in hours with pre-made software. A fake video of the CEO, a manipulated voice recording giving the green light to wire transfer or a fabricated press release can be created and sent out before any internal system of verification can react. The damage that occurs during that window is often permanent [1].

Synthetic media is a wide and growing category of AI-generated or manipulated video, voice, text and images. Initial academic and policy interest was mostly directed towards political disinformation [2]; the corporate sector is an emerging primary target. There are real monetary costs and benefits. The damage that can be caused by a manipulated video suggesting product contamination, a false audio instruction instructing a large transfer, or a completely bogus earnings statement can each be measurable within hours of it being circulated [3]. These risks are not limited to the economies in the west. Despite having the largest number of internet users in the world (more than 850 million) and a fast growth of the corporate

presence online, India remains a focus of synthetic media manipulation, and one that has been very under-researched in the academic literature on corporate deepfake risk.

In this paper, there are three special points that are set out. The first is to create a map of how to build and use synthetic media attacks against corporate entities. The second is to recognise the governance and institutional context of organisations that makes them more or less vulnerable. The third is to assess the effectiveness of existing technical detection techniques and regulatory regimes, both on a global level and in the Indian context.

2. MATERIALS AND METHODS

2.1 Study design

The objectives are pursued by systematically reviewing peer-reviewed literature published on the subject during 2019-2024, Incident Reports, Legislative texts from India, EU, USA, and UK, and Technical publications about detection methods. The sources were mainly journals that were indexed in Scopus and Web of Science, and government documents and verified journalism were added when the number of academic literature was limited. The paper is not empirical data but rather is a synthesis of the existing data to draw conclusions which can guide practice and policy.

2.2 Scope and jurisdictional focus

Here, the Indian dimension is not given as a footnote. India has a unique set of risks: a massive and expanding pool of digitally savvy corporate workers; enforcement systems still trying to catch up with risks arising from AI; a legislative framework for digital regulation that is undergoing rapid transformation; and a social media landscape where misinformation is quickly shared across WhatsApp, YouTube and regional language platforms and is frequently not detected by English-language monitoring tools. Section 3 develops the analysis in four parts attack mechanisms, governance and institutional vulnerability, detection and regulation, and recommendations before Section 4 concludes.

3. RESULTS AND DISCUSSION

3.1 Mapping synthetic media attack mechanisms in corporate environments

3.1.1 Mechanisms of attack development and implementation

The first step in grasping the vulnerability of corporations is to understand how these synthetic media attacks are really created. Technical barrier has significantly decreased. Since 2022, voice cloning technologies have been readily accessible and can produce realistic copies from just three seconds of speech for a given speaker [4]. Generating a video deepfake, which was once limited to scientists using large clusters of GPUs, has also been made possible by a new generation of consumer-grade tools that can turn a person's publicly available photos and videos into a fake video in an hour using just their target footage [5]. The material for these attacks is largely self-supplied by the targets. The companies' executives have profiles on LinkedIn, are interviewed for the record, are called out for earnings calls, and are featured in promotional videos, which are rich with training data for voice and image synthesis.

There is a consistent attacking pattern in all recorded attacks. An opponent discovers a high value target usually an executive whose voice or picture has institutional authority. Publicly available media is used to train or fine-tune a generative model. The fabricated content is then placed in a believable context: An internal communication channel, a news clip, a social media post, or an email attachment. When making these decisions, the degree of sensitivity of the situation plays a key role, including during seasons of earnings announcements, merger discussions, regulatory filings, and product releases, when institutional credibility is at stake and a hurry up decision is needed [3].

The audio forgery of a British energy company CEO in 2019 is a prime case in point. A telephone call on behalf of the chief executive of the parent company was made to the target executive and instructed him to make a transfer of around 220,000 euros into a supplier account. The success of the call was more due to the coherence of the contextual framing, the plausibility of the operational norms, and the urgency of not verifying the synthesis instinct [6]. In the United Arab Emirates, a financial institution authorized a 35-million-dollar transfer, but was fooled by a deepfake video call that impersonated a company's legitimate director in 2020 [7].

There are documented cases in India. In 2023, a fake video of Ratan Tata, one of the most recognizable business entities in India, endorsed a fake investment scheme. The video was shared extensively on social media platforms like WhatsApp and YouTube, which, prior to the corrections, were used to raise funds from retail investors who believed the name behind it [8]. In the same year, another fake video of Mukesh Ambani was created and shared in a similar scam, featuring Ambani approving a stock trading platform that was not connected to him or his firm [9]. In both examples, Indian's high mobile internet usage, a population of first-generation digital users and low media literacy has provided fertile ground for corporate deepfake fraud.

The attacks on products are carried out via fake videos, attacks on finances are made via fake analyst briefings, and attacks on the actual press releases are conducted via fake press releases [10]. They are all victims of the same crime the exploitation of institutional trust, the assumption that the communication being offered from a known entity has a fundamental presumption of authenticity which can be undermined by the systematic exploitation of a deepfake.

3.1.2 Financial and reputational injury to corporate targets

Synthetic media attacks have measurable impacts, which can be divided into three categories: financial, market disruption, and reputational. In 2022, FBI reports indicate that business email compromises (BEC) are growing in scope and are often supplemented by synthetic audio, causing losses of more than 2.7 billion US dollars. Between 2022 and 2023, the number of complaints of financial fraud being committed using AI tools in India rose considerably, and the language and visual scams were identified as identifiable subcategories of the reported frauds in the country's financial data by India's National Cyber Crime Reporting Portal and the Ministry of Home Affairs, respectively [11].

Table 1. Documented synthetic media incidents involving corporate targets, 2019-2023

Year	Target / Country	Attack method	Reported impact	Source
2019	Energy CEO, UK	AI voice cloning phone call	USD ~243,000 transferred	Stupp, 2019 [6]
2020	Financial firm, UAE	Deepfake video in live call	USD 35 million transferred	AFP, 2021 [7]
2022	Multiple firms, USA	Synthetic audio in BEC calls	USD 2.7 bn sector total	FBI IC3, 2023 [12]
2023	Ratan Tata deepfake, India	Fabricated endorsement video on WhatsApp / YouTube	Investor fraud; scale undisclosed	ET Bureau, 2023 [8]
2023	Mukesh Ambani deepfake, India	Fake stock platform endorsement video	Retail investor losses reported	Hindustan Times, 2023 [9]

Source: Compiled by the author

Table 1 is not all-inclusive and reflects only a few of the incidents documented. The Indian cases are also included with the UK, UAE and USA cases, pointing to the worldwide spread of synthetic media corporate fraud, which hasn't been well captured in academic work. While the structural characteristics of the Indian cases are similar to their Western counterparts, the distribution channel is different: Cases on WhatsApp are more difficult to monitor and eradicate than those on public indexed platforms due to the encrypted design of the platform.

Disruption is a type of harm that is unique to market. A document using a listed company's investor relations function that is fabricated, or a persuasive video of a CEO, can lead to share price changes before a correction is released. Coordinated disinformation can cause market volatility in equity markets, even in the absence of high-tech synthetic media [13]. The credibility bar has been raised for fake content, which is more likely to be generated using AI tools, before the fact-checkers can get involved. This risk is high in India's retail-investor-dominated equity market where many investors get financial advice from social media.

Long-term reputational damage is more difficult to measure and therefore, even more important. Consistent with this, research on the psychology of misinformation has shown that even when correcting explicit belief successfully, there are lingering traces of the misinformation that lead to continued doubts about it, called the continued influence effect [14]. If the brands involved are consumer-facing, this taint can live beyond the time of debunking and actually be present in the official narratives. It's not likely that a consumer who has watched a video of an alleged product contamination incident will forget it after learning that the video is synthetic.

3.2 Governance failures and institutional vulnerability

3.2.1 Structural deficiencies and organisational risk exposure

There are at least three governance failures that affect institutional susceptibility: cyber security is separated from communications functions, synthetic media is missing from institutional crisis management plans, and the institutional culture ignores the concept of verification as a tool, and sees it as friction. For most larger organizations, information security exists within a technical unit that has little integration with communications, public relations or the executive functions. In a synthetic media attack, the goal is not to target data systems, but rather a reputation, which means it goes beyond the security perimeter without going through any systems designed to assess authenticity. A similar phenomenon was identified by Chesney and Citron [2] that simply having deepfake technology raises ambient doubt on content authenticity and diminishes the evidentiary value of authentic recordings, with organizations remaining silent on both sides of this issue.

The second failure is due to a process. The benefits of having a pre-planned response protocol to limit decision making errors in the heat of the moment have been with respect to crisis communication scholarship [15]. However, most existing corporate crisis plans prior to 2022 didn't consider synthetic media as a threat category. In 2023, Pew Research Center found that just one-third of organisational communications professionals said they had any sort of formal guidance to follow when confronted with a deepfake, and less than 15 percent had conducted simulation exercises focused on AI-generated disinformation [16]. The numbers are representative of the Western corporate world with equivalent figures for Indian organisations not being collected systematically; however, industry consultations indicate that the gap in preparedness is at least as significant.

The third failure is cultural. It takes time to verify and organisations that are focused on speed and decisiveness see verification as a delay. A recent case of audio impersonation in 2019 provides an example: The executive who was targeted said that he was under pressure to make a decision on a request from a seemingly authoritative figure [6]. The social cost of verification is higher in India due to hierarchical organizational cultures where instructions given by seniors are less likely to be challenged, leading to higher probability of fabricated authority instructions being executed without challenge [17]. AI adoption is itself reshaping organisational processes, and research in the information technology sector shows that artificial intelligence measurably affects employee performance and workplace dynamics [18].

3.2.2 Sector-level risk variation

Some organizations are more vulnerable to exposure than others. The attack surface depends on the sector, how prominent key participants are on the public radar, how much decisions can be made via verbal or video command and how much media coverage exists. Healthcare, retail and certain listed technology firms are all exposed, and the risk is far greater for pharmaceutical companies and for entities engaged in contentious public debate. FMCG giants whose founder is a big name and fintech businesses where the credibility and trust are key due to their brand names are at a higher risk in India as their leadership is well-documented in regional and national media in various languages.

Table 2. Sector risk profile for synthetic media corporate attacks

Sector	Executive exposure	Financial risk	Reputational risk	India-specific amplifier
Financial services	High	Very high	High	WhatsApp-based fraud; low investor media literacy
Technology (listed)	Very high	High	Very high	Founder personality cults; large retail investor base
Pharmaceuticals	Medium	High	Very high	Product safety rumours spread via regional platforms
FMCG / consumer goods	Medium	Medium	High	Deep rural digital penetration; Hindi and regional language deepfakes
Fintech / crypto	High	Very high	Very high	Unregulated endorsement culture; retail exposure

Source: Compiled by the author

The risk amplifiers used in Table 2 are specific to India and not included in many of the international analyses of corporate deepfake vulnerabilities. One aspect of the WhatsApp dimension is of special interest. In WhatsApp groups, the material is not indexed by the normal media monitoring tools, does not feed the automatic fact-checking tools and is consumed by users within a social environment that gives further

credibility to the content, since it seems to come from a trusted personal network. The same corporate content can be widely shared in group environments of WhatsApp and circulate for days or weeks until identified and corrected by a monitoring system [19].

The smaller the organisation, the less protected and in some ways more vulnerable they can be. They usually do not have the internal means to maintain detection systems, to carry out proactive media monitoring or have their own specialist legal representation to facilitate swift takedowns [20]. One of the structurally destabilising characteristics of synthetic media as a threat category is that it is so easy to produce an attack, but so difficult to launch an organisational response. This imbalance is starker in the mid-market corporate world in India where companies can be very well known but they do not have internal capacity to deal with digital threats.

3.3 Detection capabilities and legal responses

3.3.1 The technical detection problem

Since at least 2018, when DARPA's Media Forensics programme started funding systematic research on detection algorithms [21], substantial investment has been made in technical detection of synthetic media. The prevailing architecture is to train classifiers on paired sets of authentic and manipulated content and detect statistical artefacts that have been added by the process of content creation. Initial solutions were based on discrepancy in the visual components, such as unnatural blinking, mismatching lights, skin texture anomalies at the boundaries. More recent approaches involve neural network architectures to recognize frequency domain signatures associated with generative models [22].

The basic problem is that detection is in an ongoing arms race with generation technology, with detection generally losing. Wang et al [23] showed that classifiers trained to detect content generated by one type of generative model struggle to detect content generated by other types of models that they did not train on [23]. This is a cross-generator generalisation failure that requires ongoing retraining of systems as new tools become available in the realm of generators, which is not something that most organisations can afford. For high-quality synthetic material, Groh et al. [24] report that unaided human detection was essentially at random, and that tool-assisted detection has a considerable false negative rate in high-stakes verification situations [24].

Another issue specific to India is that media production of synthetic media in India is done in multiple languages and multiple modalities. Detection algorithms trained largely on English-language content and Western facial features work less well when the content involves South Asian faces, regional language audio files, and when the content is distributed via platforms that compress media files heavily, which embeds the artefacts that detection algorithms are trained to detect [25]. This isn't just a technical tidbit but it indicates that the detection mechanisms used by Indian organisations and regulators are not as effective as the detection mechanisms they would use in their environment where they were created.

Alternative strategies that have emerged are based on 'provenance'. The Coalition for Content Provenance and Authenticity (C2PA) is a joint cross-industry effort by Adobe, Microsoft, Intel, BBC and others that has created an open standard for content credentials that captures verifiable information about the content's creation and editing history [26]. The adoption has been uneven across the globe, and very low in India, where the content provenance infrastructure of the services in the platform ecosystem does not include C2PA or any other standards.

3.3.2 Regulatory responses across jurisdictions, including India

Since 2022, legislative attention to synthetic media has increased. As part of the EU's AI Act (2024), specific deepfake applications are considered 'high-risk' AI uses and must be labelled as such as synthetic media reaches the public [27]. In the U.S., the situation is a bit murkier: there are a handful of states with laws specific to certain contexts involving deepfakes, but no single federal legislation that controls synthetic media manipulation for commercial or corporate uses [2]. Ensuring the removal of harmful deepfake content was a key priority in the United Kingdom's Online Safety Act 2023, which has a far wider application than to corporate harm alone [28].

Indian laws relating to the harm caused by synthetic media are undergoing a dynamic transformation. The Information Technology Act 2000 and its amendments are the basis legal documents for combating online fraud and impersonation, but they were not created with the use of AI-generated content in mind, and they do not specifically address the evidentiary or procedural challenges that arise in deepfake cases [29]. Inconsistent enforcement of obligations under the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules 2021 to remove harmful content within specified timelines or set up grievance redress mechanisms has not been observed in the case of synthetic media [30].

The most notable development in the Indian law in recent times is the Digital Personal Data Protection Act, 2023, which lays down the framework for the rights of data principals and the obligations of data fiduciaries, which in theory can be applied to the biometric and behavioural information that can be used to generate deepfakes of identifiable individuals [31]. The rules for implementing the Act, however, as at mid-2024, had not been finalised, and there is no specific reference to synthetic media or AI-generated impersonation as a harm classification under the Act. The Press Information Bureau has already issued some advisories regarding deepfakes and the Ministry of Electronics and Information Technology gave a formal advisory to social media platforms to label AI-generated content and remove deepfakes within 36 hours of reporting [32]. This is not a law but rather an expression to date of the Indian government's expectation of platform responsibility for synthetic media damage.

Corporate-wise, none of these frameworks offers the best answer to what is most important in a synthetic media attack a quick and legally binding way to remove fake media before it can cause harm. The 36-hour window advisory by the Indian authorities is all but insignificant if the deepfake video goes viral in WhatsApp in the first 2 hours. The civil remedies under defamation law involve a long and time-consuming procedure to prove publication, falsity and damage, which can take months to years. The procedural challenges posed by deepfake content, such as determining to a legal standard when the deepfake is fabricated, along with the known error rates of detection tools, further complicates the existing litigation framework.

3.4 Recommendations for governance, practice and policy

The sophistication of synthetic media attacks and the lack of readiness of most corporates institutions call for parallel actions, both in governance, practice and policy. These recommendations are aimed at three groups first, corporate leadership and governance bodies, second, communications and security practitioners, and third, policymakers and focus on the Indian context where the threat/institutional gap is especially high.

The greatest demand for corporate governance is to integrate synthetic media risk as a specified, distinct category of enterprise risk. This involves carrying out an exposure audit, and determining which executives have enough visibility in the public media to be seen as a possible target of impersonation, what operational processes could be manipulated by audio or video instructions, and which product or regulatory environments could make the organisation a possible target for defamation through synthetic content. These governance duties extend beyond synthetic media: scholarship on algorithmic agents argues that institutions must confront the overt and covert challenges that autonomous AI systems pose to justice and the future of work [33]. Considering the fact that in most of the Indian deepfakes, the founders and chairman of the Indian conglomerates or listed companies have been impersonated, Indian listed companies and conglomerates with recognized leaders at the helm should consider synthetic media risk as an executive protection as well as a communications challenge.

The most significant change at the practitioner level is the addition of out-of-band verification protocols for high-value instructions. All instructions which authorize a financial transaction, material disclosure or any personnel action or any significant operational decision that exceeds certain thresholds should be confirmed through a different channel from the one used for the original instructions. The protocol should specifically cover WhatsApp, which is so popular that people at the highest levels in Indian organisations use it to communicate with each other internally and are clearly exposed to voice and video impersonation in that channel. Language monitoring should be included in media surveillance briefs for the region, as fake information aimed at Indian companies often is available first in regional languages (Hindi, Telugu, Tamil, Bengali, etc.) and then in English.

Indian policy makers, in particular, should refrain from the advisory framework and adopt mandatory precautions for platforms to detect and flag the content that is meant for high reach. The implementing rules to the Digital Personal Data Protection Act should clearly define biometric and voice data as sensitive personal data, and provide obligations for their collection and processing. The most egregious deficiency in the existing system would be provided by a fast-track judicial process for the corporate victims of these synthetic media attacks similar to the expert injunctive relief in IP cases. The international arena, including the creation of global regulations for the governance of AI-generated content, will see India benefiting from its presence by promoting through its participation in these discussions the adoption of criteria and training sets that recognize the technical bias issue mentioned earlier and include South Asian languages and facial features.

4. CONCLUSION

Most organizations are dealing with synthetic media manipulation as a current reality of their operations, but they are largely ill-equipped to do so with the proper tools, processes or institutional awareness. The evidence surveyed in this paper indicates a consistent and worrisome pattern. Attacks are easily available and inexpensive to execute. Organizations are under-resourced at governance, communication and security at the same time. The detection technology has not caught up to the generation capabilities. Regulatory structures are weak, slow to act and are broken up. The direct cost of successful attacks, from financial damages to disruptions to the market or damage to reputation, can be significant and is increasing rapidly, and the impact of these attacks can last long after the fabrication is disclosed.

This Indian aspect of the problem is not well represented in the literature and this paper has emphasized the need for special studies on it. The digital penetration rate is high, social media is a key source of information for many retail investors, the information flow is speeded up via WhatsApp, and the regulatory environment is in flux but hasn't yet caught up to the unique challenge of synthetic media corporate fraud.

The main point is that the synthetic media manipulation risk category is not the same as the type of communication incident. Structural risks must be met through institutionalised means, rooted in governance, policy and culture. A crisis communication meeting is the first step after a deep fake attack that an organisation will take, and that is the advantage it has that makes deep fake attacks work. The right posture for this involves constant adaptation, based on a realistic evaluation of current exposure; this paper has tried to offer some analytical basis for realistic assessment of current exposure in various jurisdictions, particularly in the context of India's risk profile.

ACKNOWLEDGEMENT

None

REFERENCES

[1] 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), 1-13. https://doi.org/10.1177/2056305120903408

[2] Chesney, R., & Citron, D. K. (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

[3] Bateman, J. (2020). Deepfakes and synthetic media in the financial system: Assessing threat scenarios. Carnegie Endowment for International Peace.

[4] Nautsch, A., Jimenez, A., Treiber, A., Kolberg, J., Jasserand, C., Kindt, E., Delgado, H., Todisco, M., Hmani, M. A., Mtibaa, A., Hauber, M. A., Sarkar, A., Yi, Z., Hua, G., Bonastre, J., Busch, C., & Evans, N. (2022). Preserving privacy with fine-grained voice anonymization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, 2367-2380. https://doi.org/10.1109/TASLP.2022.3190741

[5] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64, 131-148. https://doi.org/10.1016/j.inffus.2020.06.014

[6] Stupp, C. (2019, August 30). Fraudsters used AI to mimic CEO's voice in unusual cybercrime case. *The Wall Street Journal*. https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

[7] Agence France-Presse. (2021, October 17). Deepfake video call used in $35m bank heist, FBI warns. *The Guardian*. https://www.theguardian.com/technology/2021/oct/17/deepfake-video-fraud

[8] Economic Times Bureau. (2023, October 12). Deepfake video of Ratan Tata used to promote fraudulent investment scheme. *The Economic Times*. https://economictimes.indiatimes.com/tech/technology/deepfake-video-of-ratan-tata-used-to-promote-fraudulent-investment-scheme/articleshow/104374621.cms

[9] Hindustan Times. (2023, November 8). Fake video of Mukesh Ambani used to lure investors into fraudulent trading app. *Hindustan Times*. https://www.hindustantimes.com/technology/fake-video-mukesh-ambani-deepfake-trading-fraud-101699421553436.html

[10] de Ruiter, A. (2021). The distinct wrong of deepfakes. *Philosophy and Technology*, 34(4), 1311-1332. https://doi.org/10.1007/s13347-021-00459-2

[11] Ministry of Home Affairs, Government of India. (2023). Annual report on cybercrime in India 2022-23. National Crime Records Bureau.

[12] Federal Bureau of Investigation Internet Crime Complaint Center. (2023). Internet crime report 2022. FBI. https://www.ic3.gov/Media/PDF/AnnualReport/2022_IC3Report.pdf

[13] Kertysova, K. (2018). Artificial intelligence and disinformation. *Security and Human Rights*, 29(1-4), 55-81. https://doi.org/10.1163/18750230-02901005

[14] Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3), 106-131. https://doi.org/10.1177/1529100612451018

[15] Coombs, W. T. (2015). Ongoing crisis communication: Planning, managing, and responding (4th ed.). SAGE Publications.

[16] Pew Research Center. (2023). Deepfakes and digital disinformation: Public awareness and organisational readiness. Pew Research Center.

[17] Kakar, S. (1978). The inner world: A psycho-analytic study of childhood and society in India. Oxford University Press.

[18] Lalitha Kumari, P., & Pandey, J. K. (2021). Impact of artificial intelligence on employees' performance in information technology sector. *AIP Conference Proceedings, 2587*(1). https://doi.org/10.1063/5.0150403

[19] Banaji, S., Bhat, R., Agarwal, A., Passanha, N., & Sadhana Pravin, M. (2019). WhatsApp vigilantes: An exploration of citizen reception and forwarding of WhatsApp misinformation linked to mob violence in India. London School of Economics and Political Science.

[20] Chesney, R., & Citron, D. K. (2023). Deepfakes and the new disinformation war. *Foreign Affairs*, 98(1), 147-155.

[21] Guo, H., Hu, S., Xie, X., & Li, S. (2021). Fake face detection methods: Can they be generalized? In 2021 IEEE International Joint Conference on Biometrics (IJCB) (pp. 1-8). IEEE. https://doi.org/10.1109/IJCB52358.2021.9484361

[22] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Niesner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1-11). IEEE. https://doi.org/10.1109/ICCV.2019.00009

[23] Wang, S. Y., Wang, O., Zhang, R., Owens, A., & Efros, A. A. (2020). CNN-generated images are surprisingly easy to spot…for now. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8695-8704). IEEE. https://doi.org/10.1109/CVPR42600.2020.00872

[24] Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2022). Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, 119(1), e2110013119. https://doi.org/10.1073/pnas.2110013119

[25] Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Surveys*, 54(1), 1-41. https://doi.org/10.1145/3425780

[26] Coalition for Content Provenance and Authenticity. (2022). C2PA technical specification version 1.2. C2PA. https://c2pa.org/specifications/specifications/1.2/specs/C2PA_Specification.html

[27] European Parliament. (2024). Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*.

[28] UK Parliament. (2023). Online Safety Act 2023. UK Legislation. https://www.legislation.gov.uk/ukpga/2023/50/contents

[29] Ministry of Law and Justice, Government of India. (2000). The Information Technology Act, 2000 (Act No. 21 of 2000). Gazette of India.

[30] Ministry of Electronics and Information Technology, Government of India. (2021). Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021. Gazette of India.

[31] Ministry of Law and Justice, Government of India. (2023). The Digital Personal Data Protection Act, 2023 (Act No. 22 of 2023). Gazette of India.

[32] Ministry of Electronics and Information Technology, Government of India. (2023, November 1). Advisory to social media intermediaries on deepfakes and AI-generated content. MeitY. https://www.meity.gov.in/writereaddata/files/advisory_on_deepfakes.pdf

[33] Pandey, J. K., & Kumar, R. (2025). Governing the algorithmic agent: Confronting overt and covert challenges to justice and the future of work. *International Journal of Law Management & Humanities, 8*(4), 1974-1984. https://doij.org/10.10000/IJLMH.1110648

