World Research of Business Administration Journal
Vol. 6, No.2, 2026

Deepfake Threats in Business Communications: Detection, Governance, and Trust Management

https://doi.org/10.56830/WRBA07202601

Sally M Elsayed
Assistant Professor, American University in Bahrain, Bahrain
Email: sally.elsayed@aubh.edu.bh
Received March 15, 2026 Accepted June 18, 2026 Published July 25, 2026

Abstract

Deepfake technologies can create seamless impersonations of real people, making them say or do things they never actually said or did. These impersonations can be entertaining parodies, but they pose challenges for organizations. Because the technology enables smooth, easy fake video generation, it is contributing to the proliferation of fake digital media and fake news. The implications of Deepfake technologies are far-reaching, threatening democracy, physical privacy, personal reputation, and financial stability. Domestically, the challenge of managing trust is already daunting. It will worsen substantially as malicious actors continue to leverage generative models to produce increasingly convincing synthetic content. Creating strong governance frameworks is a necessary and appropriate response to Deepfakes, though it is not clear who should take the lead and how. Such fundamental tenets of modern society are core to stakeholder confidence in cross-border investment, financial systems, digital currencies, and currency exchange. As AI-based Deepfake technologies advance and proliferate, their risks to digital communications and governance systems are expected to increase. Organizations should proactively understand their influence trends and implement robust detection and evaluation methods to correct the effects of Deepfake technologies. The detection challenge spans multiple scopes and levels, extending to online comments and digital platforms. Mechanisms deployed in understanding these trends should extend beyond the corporate premise. These may involve synthetic media and Deepfake techniques, fast-evolving tools that should be anticipated and analyzed thoroughly. Such an understanding might help leverage the associated protocols before they inflict damage and may also help build competitive intelligence.

Keywords: Deepfake Threats, Business Communication, Governance, Detection, Trust Management

1. Introduction

Deepfakes—synthetic or manipulated media, particularly in the form of videos—have emerged as both a significant risk and a notable opportunity in the realm of business communications. These advanced technologies can pose serious threats to corporate reputations by generating Deepfake videos that depict executives misusing their public image, potentially leading to misinformation and public distrust. However, on the flip side, they can also serve as powerful tools for businesses to effectively communicate solutions or even counter-narratives in situations that are commercially sensitive or challenging. The implications of deepfakes necessitate thorough frameworks for

Print ISSN 2771-1161
Online ISSN 2771-1153

governance, detection, and trust management, which warrant careful consideration and should be implemented in a timely and widespread manner across all levels of the organization.

Deepfake creation and detection techniques increasingly rely on deep learning (Thi Nguyen, et al., 2019). Detecting deepfakes based on technical indicators is challenging because well-crafted videos can evade leading detection methods. Awareness of contextual and behavioural cues applied in conjunction with other signals can help people recognize potential deepfakes, even when situational specifics remain out of reach. Stakeholders may be better supported and companies' reputations reinforced by communicating up-to-date information about the deepfake landscape, articulation of deepfake risks, clarification of governance processes for managing such threats, and avenues for verification of vendor and partner communications (Kietzmann, Lee, McCarthy, & Kietzmann, 2019).

2. The Landscape of Deepfake Technologies

The rapid growth of deepfake technology represents a substantial threat to enterprises. Deepfakes employ sophisticated algorithms to generate hyper-realistic video and audio simulations of individuals by synthesizing what they say and how they behave. Initially limited to entertainment, parodies, and social misuse of celebrities, deepfake technology has now spread to all-global communications, including business and society. Deepfakes divert attention away from real issues through manipulation of trust and credibility, with important implications for governance and trust management, respectively. (Gupta & Ozen, 2025)

The threats posed by deepfake technology have intensified as access to sophisticated AI tools has widened. Corporate executives, thought-leaders in society, and their organizations are now exposed to deepfake creation that can compromise communication in varying degrees—ranging from impersonation through imitation to establishing credibility and influence through behavioral mimicry. Personnel stability and mental safety have been jeopardized, resulting in adverse effects on employees, resources, and firm strategy. (Singh, 2023).

3. Detection and Verification in Corporate Settings

Deepfake detection solutions can be categorized into technical detection methods that involve algorithms and the incorporation of behavioral and contextual cues that assist in the authentication process. Technical methods are generally based on trained algorithms to distinguish between real and manipulated content yet they are time-consuming to develop and constantly need updates to adapt to a fast-evolving threat. Furthermore, they often generate false positives which may render them unsuitable for corporate use (Kietzmann, Lee, McCarthy, & Kietzmann, 2019); (D. Bray, D. Johnson, & Kleinberg, 2022).

Corporate settings widely use behavioral and contextual cues to frame users' assessments of content authenticity. Only 2.5% of a corporate deepfake training set was evaluated using these approaches. The public organization employed mainly syntactically based fake detection mechanisms derived from public datasets. Five mandatory validation conditions were outlined for each potential data application, yet it was proved to be infeasible to comply with all conditions for their video datasets. (Heidari, Jafari Navimipour, Dag, & Unal, 2024).

3.1. Technical Detection Methods

Deepfake detection research has grown rapidly since 2017, with numerous technical approaches introduced over the past few years. Nevertheless, their performance remains limited against current

deepfake models. Many existing solutions rely on low-level artifacts that depend on specific generation methods, serving as detection points. As deepfake techniques evolve, such artifacts may become irrelevant for future countermeasures. Some works recommend combining artifact-based methods with high-level consistent properties like head pose or facial expression to enhance robustness. (Almutairi & Elgibreen, 2022).

However, stable detection remains challenging, as latent spaces can still capture and transport coarse artifacts between real and fake samples, making the combination strategy difficult to implement. Furthermore, the vast number of deepfake generation algorithms hinders research focused only on detection (D. Bray, D. Johnson, & Kleinberg, 2022).

In addition, many techniques target a single generation method and thus offer limited utility for upstream detection solutions, leaving a pressing need for more generic, effective methodologies capable of countering a wide range of approaches. Models either duplicate expansions across both fake and real domains or leverage a pre-trained generator to re-produce additional fake and real samples. Instead of generating fake samples from cleaned, source-target matched images for single-algorithm detection, attention could be directed toward acquiring a broader dataset containing diverse real and fake samples across multiple widely adopted methods to facilitate simultaneous detection of various popular algorithms. (Armingol, Baghdassarian, & Lewis, 2024)

Empirical observations reveal that the difficulty of distinguishing between deepfake videos generated by state-of-the-art (SOTA) algorithms varies significantly depending on the specific SOTA technique used to conduct generation. Although noticeable quality differences exist among diverse SOTA samples, an overwhelming share of existing detection solutions prioritises performance on a narrow range of commonly employed generation techniques. Remaining uncertainties in the generative architecture of diverse deepfake models leave ample opportunity to develop detection methodologies for these generating schemes that enhance the overall robustness and universality of existing SOTA counter-fakes. (Edwards, Nebel, Greenhill, & Liang, 2024).

3.2. Behavioral and Contextual Cues

Content received and processed correctly. Here's the requested section.

A second line of defense involves the use of behavioral and contextual cues to assess information (Kietzmann, Lee, McCarthy, & Kietzmann, 2019). Deepfake detection technologies have steadily advanced. However, a general assurance that deepfake videos, audio recordings, or images are safe cannot be given. Deepfakes can be detected through technical analysis of attributes such as audio frequency, visual patterns in a face, or heartbeat synchronization that typically occur in human interactions.

The effectiveness of detection tools varies with different types of deepfake materials. Audio or video recordings extracted from conferencing platforms or social media may exclude important signatures that are detectable in material generated by conventional applications. As a consequence, these items might escape detection. Attackers are known to generate deepfakes in a manner that evades machine detection. Deepfake software may be adapted for purposes unrelated to deepfake generation, further complicating technical detection. (Heidari, Jafari Navimipour, Dag, & Unal, 2024).

Print ISSN 2771-1161
Online ISSN 2771-1153

3.3. Limitations and False Positives

Deepfake systems can produce audio or video content that misleads viewers about specific elements, such as when a person appears to express a thought or feeling that is contrary to what they actually think or feel. As a result of such misleading content and the inability of technical countermeasures to detect the falsity of deepfakes, detection systems may yield a high number of false positives—they may identify an authentic communication as a deepfake. In interviews about artificial intelligence, for example, the interviewed celebrity misinterprets a question or tries to deflect it while speaking hesitantly or in monosyllables. Detection systems can interpret these signals as indications of a deepfake, leading to false positives. Such systems yield not only concerns about their robustness in dealing with highly variable and ambiguous content, but also suggestions for counteracting the effects of deepfake generation (Kietzmann, Lee, McCarthy, & Kietzmann, 2019); (D. Bray, D. Johnson, & Kleinberg, 2022).

4. Governance Frameworks for Deepfake Risk

Deepfakes present a range of significant risks that encompass various malicious acts, propaganda efforts, election interference, and widespread disinformation, all of which threaten the very foundation of democratic governance. The potential for related misuses extends far beyond the political realm, affecting corporate communications, damaging brand integrity, enabling fraud, and facilitating acts of sabotage. It is imperative for organizations to prioritize the development of robust policies and ensure compliance with regulations that address deepfake-related risks. To aid in this endeavor, governance frameworks provided by the European Union and the U.S. National Institute for Standards and Technology (NIST) offer crucial and relevant guidance. NIST emphasizes the importance of organizational preparedness by systematically defining various risks, clarifying roles, and establishing clear incident-response procedures. In addition, NIST underlines the necessity of implementing effective content verification and detection procedures to mitigate the risks associated with deepfakes and to safeguard against the potential consequences they entail. (Kietzmann, Lee, McCarthy, & Kietzmann, 2019).

4.1. Policy Development and Compliance

Governance frameworks can help mitigate deepfake risks in business and corporate communications. The first step in developing a corresponding strategy is building a set of policies and guidelines that align with existing corporate governance initiatives. These documents should be reviewed regularly and continuously updated to address new gaps or evolving threats (Kietzmann, Lee, McCarthy, & Kietzmann, 2019).

Organizations typically have policies regarding digital communication, social media, travel, harassment, information security, and crisis management; therefore, deepfake consideration can be integrated into and, importantly, governed by these streams. Gaps may be identified by answering the following questions: (i) when effective deepfake detection is available, what, if any, restrictions should be applied to deepfake use within the organization, and who can or cannot deploy the technology? (ii) when detection capabilities are still developing and detection is uncertain, what restrictions should be introduced? (iii) under which circumstances does the use of deepfake technology pose a critical risk to the organization? (iv) how could the organization's overall capability to assess incoming deepfakes be enhanced without introducing new risks? Deepfake

technology could also further augment legitimate messages in a business context, enabling better personalisation or tuning their tone or style to different audiences; exploring these opportunities may also reveal relevant controls. (Arora, Kalia, & Arora, 2025).

4.2. Roles and Responsibilities

(Kietzmann, Lee, McCarthy, & Kietzmann, 2019) describe deepfakes as fabricated or manipulated content generated by deep-learning techniques. They note that deepfakes can serve malicious purposes by producing fake video and audio of individuals delivering false messages or statements.

Among the most serious risks that deepfake technology poses to society are the destabilization of democracy, violations of privacy or reputation, and financial losses. Examples of potential malicious acts include election interference, the dissemination of harmful propaganda, and spreading and fabricating misinformation that threaten the integrity and functionality of democratic governance. (Farid, 2022).

Two types of measures can help manage deepfake risks effectively. First, it is important to combine recording original content with automatic timestamping that embeds time-linked metadata directly within the content. Life-logging technologies, including sensors that follow every activity, and specific recording processes for critical information would enable capturing original, traceable events that take place anywhere and anytime. Content proven to be original would remain as evidence against faking. Second, genuine content should be authenticated through protecting the whole workflow, enabling verification of the integrity of contents and detecting any form of manipulation. A non-intrusive method involving real-time cryptographic stamping can serve as a watermark of authentication. (Seng, Mamat, Abas, & Ali, 2024).

4.3. Incident Response and Recovery Planning

Corporate and societal contexts exhibit dissimilarities that necessitate distinct methods to determine the requisite scope and nature of incident response measures. Conversely, the response to deepfake events within corporations generally lacks specificity. Because all corporate deepfake occurrences impact the same fundamental objectives—a credible brand and customer trust—incident response preparation can concentrate on safeguarding only these pillars. This focus simplifies the response framework and integrates corporate deepfake threats into broader organizational risk strategies. (Savale, Kaur, & Kamble, 2026).

Planning for incidents involving deepfake materials entails three primary actions. First, stakeholders must delineate potential circumstances that might warrant retaliation, specifying the own- and third-party triggers for each context. Second, they need to enumerate target intervention types—such as remediation measures, persuasion attempts, or transparency communications—and establish logical prerequisites for implementing each option. Third, they should identify appropriate potential responders, clarifying responsibilities based on the nature of the intervention and the affected corporate attributes. Specifying these preparatory elements allows for a unified approach to incident management across various labeling campaigns and simpler integration of deepfake threats into existing corporate risk frameworks. (Abdel-Wahab & Alkhatib, n.d)

5. Trust Management and Stakeholder Confidence

The proliferation of these technologies diminishes trust and confidence in audiovisual communications. Some stakeholders may not even engage with such materials, expressing concern

Print ISSN 2771-1161
Online ISSN 2771-1153

about data and privacy and preferring limited information-sharing instead (Kietzmann, Lee, McCarthy, & Kietzmann, 2019). Visual communications are key to stakeholder engagement, but without effective deepfake detection, their trustworthiness depends largely on the governance of video material.

Stakeholder confidence is therefore enhanced by risk communication strategies, a commitment to transparency, and accurate vendor profile and partner verification (Etienne, 2022). Communication relies on understanding how to convey the issues at stake. Transparency extends to disclosure of policies, tools, and technologies, and monitoring of networks and materials. Trust in partners and vendors can be promoted by assessing their own governance and detection.

5.1. Risk Communication Strategies

Deepfake technologies, which are designed to create hyper-realistic synthetic audio and video that convincingly represent individuals, have the ability to generate entirely fictitious dialogues and scenarios. These deceptive practices can severely harm an organisation's brand reputation, lead to a decline in financial status, or negatively impact its competitive edge in the marketplace. To combat these threats, effective risk communication strategies play a crucial role in enabling appropriate and timely responses. By employing these strategies, organisations can signal to stakeholders and the public that they are fully aware of potential threats and possess a robust governance framework that is capable of addressing such issues. This governance framework not only outlines how the organisation will respond to various crises but also ensures that responses to damaging events, such as fabricated recordings that portray a CEO engaging in fraudulent business transactions or suggesting sexual misconduct, are handled professionally and effectively. (Kietzmann, Lee, McCarthy, & Kietzmann, 2019).

5.2. Transparency, Disclosure, and Reputation

Smart contracts that are implemented on open and permissionless distributed networks are simultaneously raising a diverse range of questions about their management and control mechanisms. This situation is particularly noteworthy and significant because transactions are initiated automatically without any need for human intermediation involved at any stage. Furthermore, the consensus on the validity of events and transactions is reached automatically by the system itself, which can lead to various far-reaching implications for governance and control within these complex technological environments. As we continue to explore the potential of these smart contracts, it becomes evident that understanding their governance frameworks is crucial for navigating future challenges and opportunities that may arise in this rapidly evolving landscape. (Wahab, Wang, Shojaei, & Ma, 2023).

5.3. Vendor and Partner Verification

Verifying the authenticity of third-party communications is critical to trust management strategies when deepfake threats are pervasive. Although most organizations do not directly commission deepfake-generated communications from external parties, indirect threat vectors, such as third-party software, add complexity to the scenario. Evaluation methods used to analyze vendors' offers may assist in assessing their susceptibility to risk, depending on vendors' trustworthiness in the marketplace and geographical location along with any historical data of communications that might already exist. While regulatory provisions concerning third-party deepfakes for sensitive content are

still emerging, consideration of controls alongside terms of engagement may warrant investigation to diminish the risks throughout the entire supply chain (Kietzmann, Lee, McCarthy, & Kietzmann, 2019).

6. Conclusion

Deepfake technologies present considerable threats to the foundational trust we have in digital communications as well as the governance frameworks designed to protect against various impersonation attacks. These threats can specifically target individuals, which may include influential leaders within businesses and key institutions, in addition to posing risks to entire organizations. Given the potentially severe implications of these technologies, a comprehensive and strategic approach is urgently needed to effectively mitigate the associated risks and bolster stakeholder confidence in digital interactions. Implementing such measures will be crucial for preserving the integrity of our communications and the trust that underpins them.

Highly public figures, especially those holding significant roles in politics, are often prime targets for the creation and dissemination of deepfakes. If attackers can convincingly impersonate these individuals, the potential for substantial harm is incredibly high. Critical decisions that leaders must make—decisions that are vital for preventing loss of life or protecting strategic interests and ensuring the prosperity of nations or organizations—could inadvertently fall into the deceptive hands of adversaries. This alarming reality raises serious concerns about security, trust, and the integrity of information, as well as the very foundations upon which democratic societies are built.

Deepfake attacks represent extremely serious threats that require a deep understanding and proactive measures to combat them effectively, especially for major organizations that play a significant role in the global economy. It is vital to recognize the potential dangers posed by these sophisticated technologies.

References:

Abdel-Wahab, A., & Alkhatib, M. (n.d). Toward Robust Deepfake Defense: A Review of Deepfake Detection and Prevention Techniques in Images. *Computers*.

Almutairi, Z., & Elgibreen, H. (2022). A review of modern audio deepfake detection methods: challenges and future directions. *Algorithms*.

Armingol, E., Baghdassarian, H. M., & Lewis, N. E. (2024). The diversification of methods for studying cell–cell interactions and communication. *Nature Reviews Genetics*.

Arora, G., Kalia, Y., & Arora, N. (2025). Sustainability in the Deepfake Era—The Role of CSR and Ethical Leadership in Mitigating Risks. In Navigating the Deepfake Conundrum:. *A Manager's Roadmap* , (pp. 49-62). Cham: Springer Nature Switzerland.

D. Bray, S., D. Johnson, S., & Kleinberg, B. (2022). Testing Human Ability To Detect Deepfake Images of Human Faces.

Edwards, P., Nebel, J. C., Greenhill, D., & Liang, X. (2024). A review of deepfake techniques: Architecture, detection, and datasets. *IEEE Access*.

Print ISSN 2771-1161
Online ISSN 2771-1153

Etienne, H. (2022). The future of online trust (and why Deepfake is advancing it).

Farid, H. (2022). Creating, using, misusing, and detecting deep fakes. *Journal of Online Trust and Safety*.

Gupta, S., & Ozen, E. (2025). of Deepfakes in G7 Business Practices. Navigating the Deepfake Conundrum: A Manager's Roadmap, 207.

Heidari, A., Jafari Navimipour, N., Dag, H., & Unal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 14(2), e1520.

Kietzmann, J., Lee, L., McCarthy, I., & Kietzmann, T. (2019). Deepfakes: trick or treat?

Savale, T. K., Kaur, D., & Kamble, S. M. (2026). Leveraging AI and Business Intelligence for Financial Decision-Making in the Age of Deepfakes. *Enhancing Consumer Trust and Marketing Resilience*.

Seng, L. K., Mamat, N., Abas, H., & Ali, W. N. (2024). AI integrity solutions for deepfake identification and prevention. *Open International Journal of Informatics*, 12(1), 35-46.

Singh, D. P. (2023). Exploding AI-generated deepfakes and misinformation. *A threat to global concern in the 21st century*.

Thi Nguyen, T., Viet Hung Nguyen, Q., Tien Nguyen, D., Thanh Nguyen, D., Huynh-The, T., Nahavandi, S., et al. (2019). Deep Learning for Deepfakes Creation and Detection. *A Survey*.

Wahab, A., Wang, J., Shojaei, A., & Ma, J. (2023). A model-based smart contracts system via blockchain technology to reduce delays and conflicts in construction management processes. *Engineering, Construction and Architectural Management*, 30(10), 5052-5072.