Trust, Identity, and Security in the Metaverse: A Survey of AI and Blockchain Perspectives

Sairathna M[1*], Sharmila Devi PK[1,] Swathi Muthukaruppan[1]and Shivakumar N[1]

[1]Department of Computer Science and Engineering, Thiagarajar College of engineering, Madurai, India
sairathnamuralidharan@gmail.com*, sharmiladevi2932006@gmail.com, swathimuthukaruppan@gmail.com, shiva@tce.edu

Abstract. The interchange of processes, systems, and users between Artificial Intelligence (AI), Blockchain, and immersive technologies such as Augmented Reality (AR) and Virtual Reality (VR) have expedited the advancements and evolution of the Metaverse into a decentralized and engaging digital ecosystem. While rationalizing and enabling intelligent and engaging immersive experiences in contexts that range from social to industrial to education, new cybersecurity incidents arise that include, but are not limited to, immersive phishing, impersonation through deepfakes, and identity spoofing. This survey reviews various AI–Blockchain frameworks for trust management, decentralized authentication, and behavioral anomaly detection in the Metaverse in the contexts of social uses, industrial uses, and educational uses. AI supports adaptive threat detection, while Blockchain supports verifiable, tamper-proof identity verification. In addition, this paper discusses challenges in regard to trust, scalability, bias, and interoperability, and provides future directions for privacy-preserving and context-aware security in the Metaverse space.

Keywords: Metaverse, Artificial Intelligence, Blockchain, Immersive Phishing, Trust Management, Decentralized Authentication.

1 Introduction

The Metaverse is being developed as a decentralized virtual environment that combines Virtual Reality (VR), Augmented Reality (AR), Extended Reality (XR), Artificial Intelligence (AI), and Blockchain technologies to facilitate immersive interactions in social, industrial, and educational contexts. The Metaverse is considered an underlying infrastructure for Web 3.0, which provides support for embodied avatars, digital assets, and collaborative interactions in shared three-dimensional spaces [14], [15]. Unlike conventional web-based platforms, immersive technologies are based on spatial computing and behavioral analysis, which increase capabilities as well as vulnerability surfaces.

The integration of these technologies also poses new cybersecurity and privacy challenges to the Internet. Research has shown UI attacks in social VR [1] and mani-

R. Vasanth Kumar Mehta et al. (eds.), *Proceedings of the International Conference on Intelligent Systems for a Sustainable Future (ISSF 2026)*, Atlantis Highlights in Intelligent Systems 16,
https://doi.org/10.2991/978-94-6239-693-7_87

pulated mixed reality content that erodes trust [2]. Phishing attacks have been transformed into immersive and AR-based smishing attacks that utilize contextual information [3], [4], which use avatar embodiment, voice conversion, and deepfakes for identity spoofing. Decentralized XR platforms also pose risks to users in terms of avatar spoofing, identity theft, and malicious virtual entities that affect psychological safety [5], [10].

To counter these challenges, AI and Blockchain have been identified as additional security solutions. AI-based solutions provide adaptive defense strategies through behavioral analysis, anomaly-based detection, and intrusion detection in immersive environments [8], [9], including VR-based learning environments [13]. Blockchain technology provides identity verification, tamper-proof authentication, secure asset validation, and data tracing in industrial and digital twin contexts [3], [6], [7], [12].

Despite existing surveys on metaverse security [11] and the individual use of AI or blockchain solutions [8], [15], most surveys consider them as distinct systems. Few surveys combine adaptive behavioral intelligence with decentralized trust verification. Immersive phishing and identity attacks are relatively unexplored from a unified AI-Blockchain solution strategy.

This survey (i) classifies immersive threats, (ii) classifies studies based on application areas, (iii) compares AI-based and blockchain-based security solutions, (iv) identifies architectural deficiencies, including scalability and interoperability, and (v) provides future research directions towards privacy-preserving and trust-resilient architectures.

2 Literature Review

Phishing in the Metaverse is a new type of social engineering that takes advantage of immersive realism, along with trust in augmented and virtual reality environments. Unlike traditional phishing, phishing in the Metaverse uses avatars, gestures, and other spatial cues to trick users. Recent advances in AI, Blockchain, and immersive environments have begun to address these threats through defenses on the spectrum of AI, Blockchain, and immersive environments. The present activity summarizes the most important studies, organized by the phishing types and defenses based on AI and/or distinct Blockchain types.

2.1 Immersive Phishing Threats in the Metaverse

Immersive phishing in the Metaverse translates social engineering into realistic, multimodal experiences. Attackers utilize avatars, gestures, voices, and virtual settings to compromise users within trusted digital environments. All of these threats can be navigated smoothly in immersive contexts themselves, establishing anonymity from traditional phishing detection techniques based on text. Table 1 outlines critical investigations that have tackled key forms of immersive phishing, specifically avatar

impersonation, voice/deepfake phishing, gestural impersonation, and environmental spoofing, in addition to their defense methods based on AI and Blockchain. AI methods focus on detecting behavioral anomalies whereas Blockchain provides a tamper-proof means of verifying identities in a decentralized way.

Table 1. Overview of literature on phishing threats and AI–Blockchain defenses in the Metaverse

Phishing Category	Representative Works (Ref.)	Key Problem Addressed	AI / ML Contribution	Blockchain Contribution	Application Domain
Avatar Impersonation	[1], [5]	Users deceived by fake or manipulated avatars in social VR	Behavioral pattern analysis and UI anomaly recognition	None	Social VR / Community Spaces
Voice & Deepfake Phishing	[8], [13]	Synthetic voices or deepfake audio impersonate trusted users	Voice biometrics, emotion recognition, deepfake detection	None	Educational / Communication VR
Gestural or Behavioral Spoofing	[2], [10]	Attackers replicate gesture or motion patterns	ML-based behavioral forensics; motion dynamics profiling	None	Mixed / Social Reality Environments
Environmental or Contextual Manipulation (Smishing / Visual)	[3], [4]	Malicious AR overlays, fake tokens, or manipulated spatial objects	Real-time anomaly detection of interactions	Anti-phishing authentication via Blockchain	AR / Economic Metaverse
Integrated AI–Blockchain Trust Frameworks	[6], [7], [9], [12], [14], [15]	Broader metaverse security: identity spoofing, intrusion, data integrity	Federated learning, anomaly modeling, multimodal fusion	Decentralized IDs (DID), smart contracts, data provenance	Industrial / Web 3.0 Metaverse

Avatar Representation. The avatars in the Metaverse share the identity of individuals, creating significant opportunities for practitioners to impersonate with phishing threats. Attackers can impersonate the avatar of a trusted user to convince others to donate their data or digital assets. Lee et al. [1] revealed this issue with their work in Illusion Worlds, where users could be misled by user interface affordance modifications by malicious actors. Likewise, Chaudhari et al. [5] considered the role of social engineering in virtual communities, showing that fake avatars could socially exploit trust networks. Because the impersonation occurs at the behavioral or visual level, and does not involve credential theft, it is difficult for users to detect this behavior.

Voice and Deepfake Phishing. Phishing has also entered the audio realm with attackers using AI-generated voice generation to impersonate legitimate users. Valluripally et al. [13] examined the impact of user immersion in audio disruptions in a

learning context and discovered there was deepfake audio that could potentially have been used for manipulation during classroom activities. Additionally, Awadallah et al. [8] classified voice-based spoofing as one of the leading threats as a result of AI, advocating for voice biometrics management and emotional signal consistency checks.

Deceptive Behavior and Gestural Spoofing. Behaviour through gesture & motion in Augmented Reality and Virtual Reality (AR/VR) environments is a way in which users are able to express their identity. This also opens up opportunities for malicious actors to tap into, either record or replicate the behavioural patterns of users and use this information to impersonate users, effectively fooling other users into thinking they are who they say they are. In the study of Kilger et al, which looked at Mixed-Reality Sensor Data, it was found that manipulation and fabrication of Mixed-Reality Sensor Data could be done continuously. Nnamonu et al. explored Gesture-based Spoofing Attacks within the Social Virtual Reality environment. To address this issue, behavior modelling using Artificial Intelligence (AI) will assist in detecting spoofing attacks based on Motion Dynamics, Timing Patterns and Deviations from expected User Responses.

Contextual Manipulation. There are also instances where the phishing attack manipulates the environment entirely, i.e including manipulating spatial objects, chat windows, or digital assets to mislead other users. Kanaoka and Isohara [4] called this "smishing" in AR. In this situation, users observed attackers project fake notifications or messages in the visual field of view. Chen et al. [3] also found a specific mode of economic phishing, utilizing fakes NFTs or virtual tokens segment to the interactivity of the environments, bridging social deception with economic deception.

2.2 AI-Driven Defenses Against Immersive Phishing

AI is a flexible defense mechanism that identifies abnormal behaviors and distinguishes between artificial or false signals. Awadallah et al. [8] noted that multimodal fusion approaches by combining facial, gesture, and voice data can be successful in combating immersive phishing. Truong and Le [9] contributed a blockchain-assisted machine learning application for decentralized intrusion detection, introducing additional privacy-aware models. Valluripally et al. [13] and Chaudhari et al. [5] explained that deep learning methods (e.g., convolutional neural network (CNN), recurrent neural network (RNN)) can detect discrepancy in emotion, gesture, or space behavior that is suspicious and indicative of spoofing behavior. However, AI systems face challenges, such as bias, privacy issues, and limited demographic diversity when developing, which calls for ethical, edge-based, and inclusive design practices.

2.3 Blockchain-Based Identity and Trust Anchors

While AI identifies attacks, Blockchain provides verifiable and tamper-resistant identity verification. Chen et al. [3] presented a Blockchain-supported anti-phishing identification protocol for the Metaverse context, which utilized elliptic curve cryptography and chaotic mapping to perform secure verification. Badruddoja et al. [12] intro-

duced the idea of using a combination of trusted AI and Blockchain to enable automatic verification of digital entities using smart contracts. Zhang et al. [14] and Fu et al. [15] articulated architectural visions in which decentralized digital identities (DIDs) would be the core of user authentication in future Web 3.0 Metaverse scenarios. DIDs would allow for trustworthy verification of avatars and non-fungible tokens, and would lead to decreased impersonation possibility. Chatterjee et al. [7] and Song et al. [6] demonstrated the ability to use Blockchain to verify digital twins and industrial assets, and therefore a similar use case exists for user identity reliability. Despite the advancement in Blockchain, these solutions can be hindered by real-time verification of latency and interoperability of verification in virtual spaces. Moreover, the space to combine AI detection and Blockchain to verify remains open to research.

3 Research Gaps and Discussion

From the above literature review, it has been observed that Artificial Intelligence and Blockchain have their respective but complementary roles in securing the Metaverse against immersive phishing attacks. The Artificial Intelligence is primarily responsible for multimodal adaptive phishing behavior detection, including user gestures, voice, and avatar interactions, whereas the Blockchain is primarily responsible for identity verification in a decentralized environment. However, most of the existing solutions are still fragmented and primarily focus on a particular type of immersive phishing attack, such as avatar phishing or deepfake phishing attacks, using a single modality.

3.1 Summary of Observations

On the basis of the fifteen representative studies surveyed [1–15], the following key observations can be made: AI-based detection systems are the most common in the current literature, with about 40% of the studies. These studies are focused on behavioral anomaly detection in immersive environments based on modalities such as social gestures, voice intonation, and avatar movements to detect phishing attacks [1, 8]. Blockchain-based trust systems account for about 35% of the studies, which are focused on decentralized identity verification and secure transaction processing [3, 6]. Only about 15% of the studies are based on combined AI and Blockchain-based systems, which indicates that the investigation of combined cross-domain defenses is still in its infancy stages [12, 14].The above ratios indicate an imbalance in the research, which highlights the fact that although AI is efficient in phishing action identification and Blockchain is efficient in trust establishment, very few systems are available that integrate these two technologies in real-time, multi-modal environments. Moreover, most of the current systems are domain-specific, which are designed for specific types of attacks such as deepfake phishing or avatar impersonation, and hence are less relevant in the heterogeneous Metaverse environment [4, 13].

3.2 Research Gaps Identified

Despite advancements, there are still a number of significant gaps in the development of phishing-resistant Metaverse solutions. Effective real-time responses to immersive phishing attacks are limited by the majority of studies that look at Blockchain-based verification and AI-based phishing detection independently [9, 15]. As illustrated in Figure 1, there is a glaring disparity and inadequate cooperative defense tactics as roughly 40% of research is devoted to AI-based anomaly detection, 35% to trust mechanisms, and only 15% to integrated AI–Blockchain approaches. Instead of providing cross-domain systems that integrate behavioral detection, decentralized identity (DID) verification, and trust evaluation, the literature is extremely fragmented and usually focuses on a single phishing modality. Table 2 outlines the main gaps, which include the inability of AI models to process real-time 3D multimodal data, the incompatibility of DID systems, the limited and biased gesture-speech-movement datasets that run the risk of incorrectly classifying users who are neurodivergent or differently abled, and scalability issues where Blockchain latency impedes VR/AR environments that need quick response times. Furthermore, the lack of standardized AI–Blockchain architectural guidelines hinders the development of comprehensive frameworks, and the lack of publicly available immersive phishing datasets limits benchmarking and reproducibility. A fully integrated, privacy-preserving, scalable, and context-aware AI–Blockchain architecture for real-time immersive security has not yet been developed, despite domain-specific advancements. This highlights the necessity for coordinated and flexible research efforts.

Table 2. Key research gaps in AI–Blockchain Metaverse security

Research Area	Identified Gap	Implication
AI-Driven Detection	AI models aren't set up to support 3D multimodal real-time data streams	Less responsive immersive environment experiences
Blockchain Integration	No operability between identity management systems using DIDs	Trust varies across multiple platforms.
Data Bias & Ethics	Limited variety of Gesture, Speech and Movement datasets	Neurodivergent or differently-abled users may be at an increased chance of being misclassified
System Scalability	Blockchain introduces additional latency on heavy transaction loads	Not suited for VR/AR applications where speed is needed
Dataset Availability	Few publicly available open datasets regarding immersive phishing.	Difficult to reproduce and benchmark results
Collaborative Defense	No consistent recommendations for AI and Blockchain architectures	Limits comprehensive phishing defense in the Metaverse area.

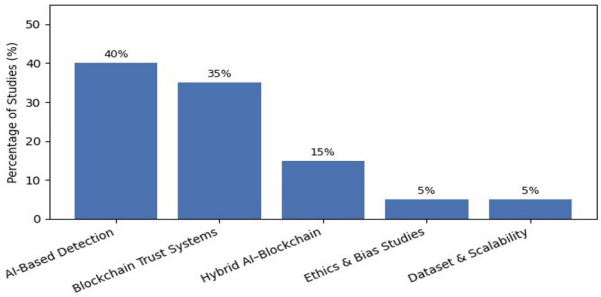

Fig.1. Overview of immersive phishing attack vectors in the Metaverse.

4 Conclusion and Future Work

This study investigated the application of Artificial Intelligence (AI) and Blockchain in protecting the Metaverse from immersive phishing attacks, such as avatar deception, deepfake voice abuse, and context manipulation. AI enables dynamic detection through multimodal behavioral analysis, and Blockchain provides decentralized and tamper-proof identity verification. However, current approaches are still fragmented, with minimal integration, and are plagued by issues like scalability, AI bias, interoperability, and the absence of holistic datasets. Our analysis emphasizes the paramount importance of developing integrated AI and Blockchain frameworks that can provide real-time and contextually aware security for XR platforms. Future studies should aim to develop holistic frameworks that integrate behavioral anomaly detection with decentralized identity management, utilizing light and low-latency Blockchain models and edge AI strategies that are privacy-centric. The creation of benchmark datasets that accurately reflect immersive interactions will also aid in this aspect. By doing so, future studies can help create secure, robust, and ethics-driven Metaverse platforms that promote trust, privacy, and inclusivity for all.

References

1. Lee, J., Heo, H., Woo, S., Kim, M., Kim, J., Kim, J.: Illusion Worlds: deceptive UI attacks in social VR. In: IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), pp. 1268–1269. IEEE, Saint Malo (2025).
2. Kilger, F., Kabil, A., Tippmann, V., Klinker, G., Pahl, M.-O.: Detecting and preventing faked mixed reality. In: IEEE International Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 399–405. IEEE, Tokyo (2021).
3. Chen, C.-M., Xiong, Z., Wu, T.-Y., Kumari, S., Alenazi, M.J.F.: Protecting virtual economies: a blockchain-based anti-phishing authentication protocol for metaverse applications. IEEE Internet Things J. 12(13), 24244–24258 (2025).
4. Kanaoka, A., Isohara, T.: Enhancing smishing detection in AR environments: cross-device solutions for seamless reality. In: IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), pp. 565–572. IEEE, Orlando (2024).

5. Chaudhari, A., et al.: Cyber security challenges in social metaverse and mitigation techniques. In: MITADTSoCiCon, pp. 1–7. IEEE, Pune (2024).
6. Song, J., Kang, Y., Song, Q., Guo, L., Jamalipour, A.: Distributed resource optimization with blockchain security for immersive digital twin in IIoT. IEEE Trans. Ind. Inf. 19(5), 7258–7267 (2023).
7. Chatterjee, P., Das, D., Rawat, D.B., Ghosh, U., Banerjee, S., Al-Numay, M.S.: Digital twins and blockchain fusion for security in metaverse-driven consumer supply chains. IEEE Trans. Consum. Electron. 70(3), 5688–5697 (2024).
8. Awadallah, A., et al.: Artificial intelligence-based cybersecurity for the metaverse: research challenges and opportunities. IEEE Commun. Surv. Tutor. 27(2), 1008–1052 (2025).
9. Truong, V.T., Le, L.B.: Security for the metaverse: blockchain and machine learning techniques for intrusion detection. IEEE Netw. 38(5), 204–212 (2024).
10. Nnamonu, O., Hammoudeh, M., Dargahi, T.: Metaverse cybersecurity threats and risks analysis: the case of virtual reality towards security testing and guidance framework. In: IEEE International Conference on Metaverse Computing, Networking and Applications (MetaCom), pp. 94–98. IEEE, Kyoto (2023).
11. AlQaruty, S., Qaruty, R.A., Hadi, S.A., Al-Tkhayneh, K.M.: A systematic literature review of security and privacy solutions for the metaverse. In: International Conference on Social Networks Analysis, Management and Security (SNAMS), pp. 154–160. IEEE, Gran Canaria (2024).
12. Badruddoja, S., Dantu, R., He, Y., Thompson, M., Salau, A., Upadhyay, K.: Trusted AI with blockchain to empower metaverse. In: Fourth International Conference on Blockchain Computing and Applications (BCCA), pp. 237–244. IEEE, San Antonio (2022).
13. Valluripally, S., et al.: Detection of security and privacy attacks disrupting user immersive experience in virtual reality learning environments. IEEE Trans. Serv. Comput. 16(4), 2559–2574 (2023).
14. Zhang, X., Min, G., Li, T., Ma, Z., Cao, X., Wang, S.: AI and blockchain empowered metaverse for Web 3.0: vision, architecture, and future directions. IEEE Commun. Mag. 61(8), 60–66 (2023).
15. Fu, Y., Li, C., Yu, F.R., Luan, T.H., Zhao, P., Liu, S.: A survey of blockchain and intelligent networking for the metaverse. IEEE Internet Things J. 10(4), 3587–3610 (2023).