Secure and Trustworthy Artificial Intelligence-extended Reality (AI-XR) for Metaverses

ADNAN QAYYUM, University of Glasgow, United Kingdom and Information Technology University, Pakistan

MUHAMMAD ATIF BUTT and HASSAN ALI, Information Technology University, Pakistan

MUHAMMAD USMAN, Glasgow Caledonian University, United Kingdom

OSAMA HALABI, Qatar University, Qatar

ALA AL-FUQAHA, Hamad Bin Khalifa University, Qatar

QAMMER H. ABBASI and MUHAMMAD ALI IMRAN, University of Glasgow, United Kingdom

JUNAID QADIR, Qatar University, Qatar

Metaverse is expected to emerge as a new paradigm for the next-generation Internet, providing fully immersive and personalized experiences to socialize, work, and play in self-sustaining and hyper-spatio-temporal virtual world(s). The advancements in different technologies such as augmented reality, virtual reality, extended reality (XR), artificial intelligence (AI), and 5G/6G communication will be the key enablers behind the realization of AI-XR metaverse applications. While AI itself has many potential applications in the aforementioned technologies (e.g., avatar generation, network optimization), ensuring the security of AI in critical applications like AI-XR metaverse applications is profoundly crucial to avoid undesirable actions that could undermine users' privacy and safety, consequently putting their lives in danger. To this end, we attempt to analyze the security, privacy, and trustworthiness aspects associated with the use of various AI techniques in AI-XR metaverse applications. Specifically, we discuss numerous such challenges and present a taxonomy of potential solutions that could be leveraged to develop secure, private, robust, and trustworthy AI-XR applications. To highlight the real implications of AI-associated adversarial threats, we designed a metaverse-specific case study and analyzed it through the adversarial lens. Finally, we elaborate upon various open issues that require further research interest from the community.

CCS Concepts: • **Security and privacy** → **Systems security**;

Additional Key Words and Phrases: Metaverse, AR, VR, XR, MR, secure ML, robust ML, trustworthy ML

1 INTRODUCTION

In the recent era, metaverse technology is rapidly emerging and there are a lot of potential applications that can benefit from these developments, such as healthcare, industry, business, and so on. While there is no single agreed-upon definition of a metaverse [35], the metaverse is a convergence of physical, augmented, and virtual reality and provides a powerfully immersive experience to users by allowing them to seamlessly interact with the real and virtual (computer-simulated) world. The term "metaverse" is a combination of two terms, i.e., "meta," which means transcending, and "universe," which refers to the physical universe and the current virtual world. This is the basic definition of the term metaverse, nevertheless, the literature shows that it does not have a unified definition [35].

Metaverse allows the creation of shared virtual space by connecting all virtual worlds through the Internet, where digital avatars (i.e., users) can communicate and interact with each other similar to the physical world. Key backbone technologies in the metaverse include **artificial intelligence (AI)** and **extended reality (XR)** that leverage different technological developments such as **virtual reality (VR)**, **augmented reality (AR)**, and **mixed reality (MR)**. In addition, similar to the current Internet, metaverse will leverage other concomitant technologies such as **information and communication technologies (ICT)**, 5G, and 6G, but metaverse will provide a qualitatively different experience to its users by enabling real-life-like 3-D experiences through the incorporation aforementioned technologies.

Metaverse allows for the digitalization of traditional brick-and-mortar institutions and businesses—it will be possible to develop virtual markets, digital lands, and digital infrastructure, which can be bought and sold using blockchain and **non-fungible tokens (NFTs)**, which are non-interchangeable units of data stored on a blockchain. Metaverse can be a game-changer in terms of the impact of its potential applications due to the greater immersion, involvement, and personalization possible due to AI-XR. This is the prime reason various corporations have shown great interest in the idealization of the metaverse and are making big bets on developing their own AI-XR-based metaverse ecosystems.

Metaverse is receiving increasing traction from numerous major tech companies worldwide such as Facebook (which is recently rebranded with the name "Meta"), Microsoft, Google, and Amazon. In addition, the widespread adoption of the metaverse is evident in the infusion of billions of dollars of investments by these companies in an attempt to achieve great technological transformation. However, despite such huge traction of the metaverse and its potential to transform existing ecosystems like healthcare, there are numerous challenges associated with the use of AI in the metaverse that may hinder their seamless adoption by end-users in the longer term. In addition, in the backdrop of recent technologically induced social issues, there is a palpable lack of trust and confidence in such technologies.

Since technology can be used both ways (i.e., for good and harm), it is vital that governments, corporations, and society at large seriously consider ethical and moral issues. There are many ethical questions about privacy, security, transparency, accountability, democracy, freedom of speech, and anonymity that technology alone cannot answer. Some specific concerns related to how AI-XR-based metaverse applications will impact humanity are: (1) How will AI-XR-based metaverse

applications impact and promote human values and human rights? How will AI-XR-based metaverse promote social welfare and not cause harm to society at large; (3) How can we regulate critical applications of metaverse like healthcare? (4) How do we align the commercial and technical imperatives of AI-XR metaverse applications with human values and promote a moral vision and character development? (5) How do we ensure that AI-XR metaverse application developers do not exploit or manipulate their users? Keeping in mind the aforementioned questions, in this article, we present an analysis of the security, privacy, and trust issues associated with the use of AI-XR in metaverse applications.

Contributions of this Article: To the best of our knowledge, this article is the first attempt towards analyzing the challenges associated with the use of different AI techniques in AI-XR metaverse applications. The comparison of this article with existing survey and review articles that are focused on analyzing security and privacy aspects of AI-XR metaverse applications is presented in Table 1. In summary, the following are the salient contributions of this article:

(1) We highlight various issues that arise with the use of AI in metaverse applications that mainly include security, privacy, and trustworthiness.
(2) We present a taxonomy of different potential solutions that can be used to realize secure, robust, safe, and trustworthy AI-XR applications.
(3) We identify various ML-based use cases across different layers of metaverse architecture and highlight several ML-associated vulnerabilities at each layer.
(4) We present a case study to highlight the real threat of AI-based vulnerabilities by considering a prospective metaverse application design scenario.
(5) We elaborate upon various open issues that require further development.

Organization of this Article: The rest of the article is organized as follows: Section 2 presents relevant background. The discussion of challenges related to security, safety, privacy, and trust is presented in Section 3. The taxonomy of different vulnerabilities associated with the use of ML in AI-XR metaverse applications is discussed in Section 4. Different potential solutions that can ensure security, privacy, safety, and trust in ML applications are discussed in Section 5. Various open issues that require further research attention are highlighted in Section 6. We conclude the article in Section 7. The organization of the article is depicted in Figure 1.

2 BACKGROUND

2.1 Metaverse: An Introduction

Before understanding the concept of the metaverse, it is very important to understand the related concepts that are described below.

— *Virtual Reality (VR):* In VR, the users achieve an immersive experience by donning a VR headset that allows them to enter into a virtual (computer-simulated) world thus completely blanking out the real world. The key objective of enabling immersion in VR is to provide high-fidelity user interaction to give him the feeling that the virtual world is real [60]. Prime examples of VR include Facebook Oculus and HTC VIVE VR headrests. VR has a wide range of applications, but a VR headset is required to enter the digital world.
— *Augmented Reality (AR):* In AR, the users obtain an immersive experience by blending the virtual (digital) and the real world and projecting digital content (text, images, and sounds) onto the real world. Unlike VR, AR can be realized without special equipment (like a headset) through the use of smartphones, implants, glasses, or contact lenses that are used to overlay digital content on top of the real world.
— *Mixed Reality (MR):* MR is a hybrid term that is used to refer to the conjunction of virtual and real worlds to produce new environments and experiences, where physical and digital

Table 1. Comparison of our Article with Existing Surveys and Review Papers that Are Focused on Analyzing Privacy and Security of AI-XR Metaverse Applications

Year	Authors	Focused Area	General Issues			ML Related Issues & Solutions					Background & Applications	Open Issues
			S	P	R	S	P	T	XAI	Ethical		
2018	Falchuk et al. [45]	Privacy issues and solutions for digital footprints in metaverse games.	√	√	×	×	×	×	×	×	≈	×
2020	Guzman et al. [38]	Analyzed privacy and security in MR from data-centric perspective.	√	√	≈	×	×	×	×	×	≈	√
2021	Ning et al. [98]	General focused survey on metaverse with partial discussion on privacy and security issues.	≈	≈	×	×	×	×	×	×	√	√
2021	Pietro et al. [40]	Discussed general privacy and security issues in metaverse applications.	√	√	×	×	×	×	×	×	≈	√
2022	Huynh et al. [68]	Discussed potential applications of AI in various metaverse applications.	≈	≈	×	×	×	×	×	×	√	≈
2022	Zhao et al. [176]	Security & privacy issues and solutions for four dimensions: communication, user information, scenario, and goods.	√	√	×	×	×	×	×	×	≈	×
2022	Wang et al. [150]	Presented general (non ML-associated) security and privacy related challenges for different metaverse applications.	√	√	√	×	×	×	×	×	√	√
	This article	ML-associated security, privacy, and trustworthiness challenges and solutions for AI-XR metaverse applications.	√	√	√	√	√	√	√	√	√	√

Legend: S → Security; P → Privacy; R → Regulatory; T → Trustworthy; √ → covered; × → Not Covered; ≈ → Partially Covered.
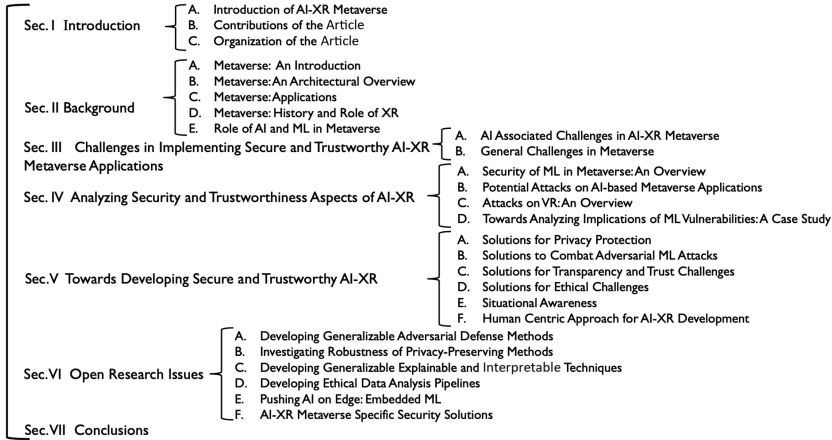
Fig. 1. Organization of the article.

objects co-exist and interact in real time (it is an enhanced form of AR). Microsoft HoloLens headrest is an example MR headset.

— *Extended Reality (XR):* XR is an umbrella term that encompasses and subsumes VR, AR, and MR. It covers all the future realities that might emerge from these technologies. XR is predicted to become a $209 billion market by 2022.

Immersive first-person experiences are one of the most significant aspects of XR, VR, and AR. The metaverse takes this to the next level, allowing large groups of individuals to share an immersive first-person experience while maintaining a strong sense of mutual presence. Although the term "metaverse" is often associated with virtual reality, according to Rosenberg, there are two types of metaverses: a "virtual metaverse," in which people are avatars, and an "augmented metaverse," in which layers of virtual content are overlaid on the real world and experienced by real people. Figure 2 depicts XR, VR, AR, and the virtual and augmented metaverse, as well as their interaction. Metaverse is the next generation of the Internet that will surround us both graphically and socially. A historical overview of developments regarding metaverse is shown in Figure 3 and different applications of XR in metaverse along with their enabling technologies are illustrated in Figure 4(a).

2.2 Metaverse: An Architectural Overview

The architecture of the metaverse expands from people's experiences to underlying enabling technologies and has seven layers that include: (1) Experience; (2) Discovery; (3) Creator; (4) Spatial
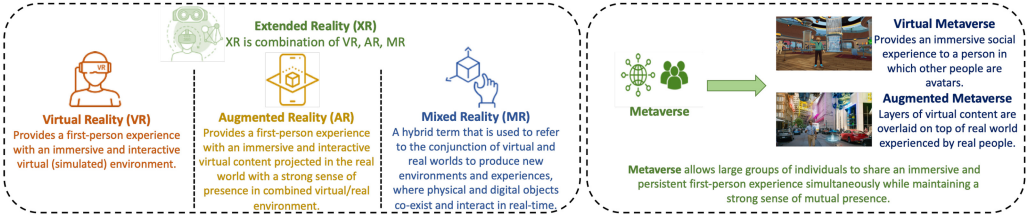
Fig. 2. An overview of different concepts related to metaverse that include VR, AR, MR, XR, virtual metaverse, and augmented metaverse.
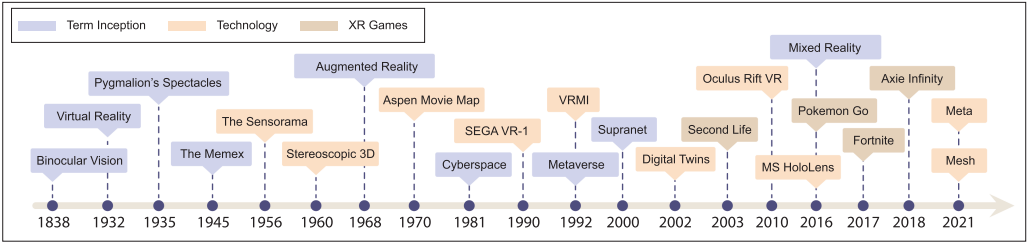
Fig. 3. Historical evolution of technologies needed for metaverse.

(a) Role of XR in metaverse.

(b) Architecture of metaverse.
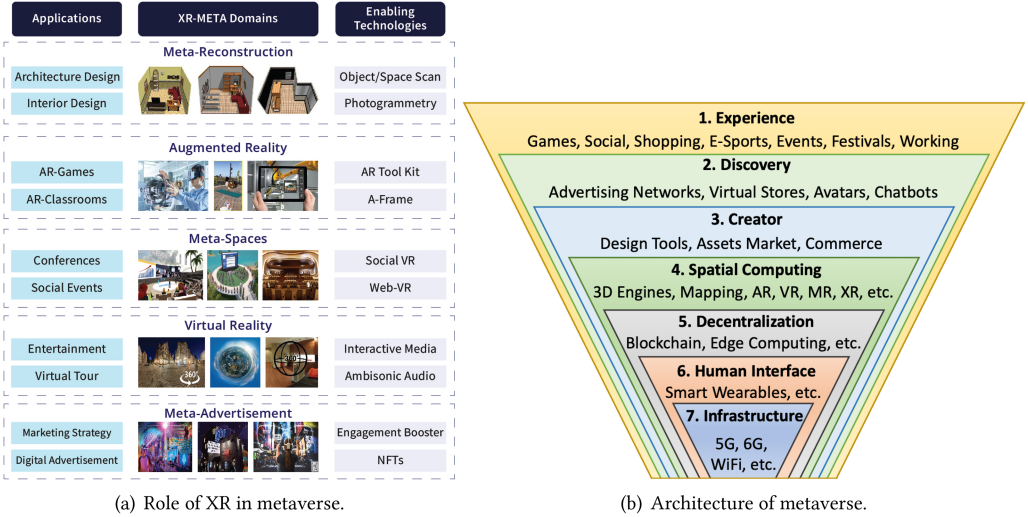
Fig. 4. Depiction of XR role in metaverse and different layers in metaverse's architecture.

Mapping; (5) Decentralization; (6) Human Interface; and (7) Infrastructure, which is illustrated in Figure 4(b) and described below.

— *Layer 1: Experience:* It is the topmost layer in the metaverse, which is mainly concerned with the experiences of the users, and it provides different services to them, e.g., games, E-sports, social interactions, events, festivals, shopping, co-working.
— *Layer 2: Discovery:* It is like a push-and-pull service that introduces people to new experiences in the metaverse, such as virtual stores, advertising networks, ratings, social curation

avatars, chatbots, and so on. It will involve both inbound (i.e., users are actively seeking information regarding experiences) and outbound (i.e., an advertisement that is not explicitly requested by the user). This layer is mainly driven by metaverse service providers and content creators to inform and motivate users regarding new features and services.

— *Layer 3: Creator:* This layer is sometimes also referred to as the creator economy. Like the previous layer, it is mainly driven by the content creator and service providers, who leverage different technologies to create content or experiences for metaverse users such as asset markets, E-commerce, design tools, and workflow.

— *Layer 4: Spatial Mapping:* This layer provides a bridge between the digital world and the physical world and provides immersive experiences to metaverse users. It consists of different technologies, such as geospatial mapping, object and speech recognition, 3D engines (for enabling animations), VR, AR, XR, multitasking, and integration of user interfaces and heterogeneous sensor data (e.g., from IoT and wearable devices). It can be assumed as the backbone of the creator layer [46].

— *Layer 5: Decentralization:* Decentralization is very crucial in the metaverse and ideally it should not be controlled by a single entity. It provides a scalable ecosystem to developers in terms of providing online capabilities and reliability to the users at the same time. This layer will consist of multiple technologies that include edge computing, blockchain, microservices, and AI agents.

— *Layer 6: Human Interface:* This layer is mainly concerned with the interfacing of the physical world with the digital and from the digital to the physical world. For example, let us consider the example of metaverse services that require data collected from humans using different sensors such as smartwatches, smartphones, smart glasses, wearable IoT devices, biosensors, and head-mounted displays, just to mention a few.

— *Layer 7: Infrastructure:* This layer is responsible for connecting different enabling devices and technologies to the network for content delivery in the metaverse. Different ICT technologies will act as a backbone in the infrastructure layer of the metaverse. For example, 5G/6G-based communication has huge potential to drastically improve the performance of metaverse applications while reducing latency and speeding up content delivery. In addition, this layer will involve major data processing capabilities such as data centers, the cloud, CPUs, GPUs, and even quantum computers.

2.3 Metaverse: Applications

Metaverse applications that incorporate different technologies such as VR, AR, or XR have various potential applications in education, healthcare, industry, and scientific research, just to name a few. A detailed taxonomy of various potential metaverse applications is illustrated in Figure 5. Metaverse allows moving from text-focused Internet that supports 2D images to a 3D or even a 4D world (in which we may travel in time (forward or backward)). One of the promising applications will be social VR, which will be the enhanced version of current social media. As metaverse can leverage both VR and AR, numerous applications (e.g., voice recognition, gesture recognition, and speech translation) can benefit [115]. Over the past few years, VR and AR technologies have become very mature and nowadays their equipment is relatively cheap and readily available. This is a long way from the modest beginning of AR and VR, which were ignited in the 1960s by Ivan Sutherland in his pioneering work on the first responsive head-mounted wearable devices, which were admittedly primitive by modern standards.

Modern VR headsets have become accessible (e.g., Facebook's Oculus Quest 2 is available for $300) with the price expected to go down as technology continues to advance. There are numerous AR applications such as **Heads-Up-Display (HUD)** features on modern luxury cars, the

Industry
- Manufacturing (i.e., developing better products and fault identification)
- Employees and Workers Training to Improve Safety
- Improving Planning for Manufacturing (Plants and Equipment)

Military
- Tactical Augmented Reality
- Immersive (Synthetic) Training Environments
- Creating physically and psychologically intensive combat situations, …

Ecommerce
- Virtual Markets and Business
- Virtual Assists
- Virtual Shopping
- AR and VR assisted immersive shopping

Education
- Improving teaching using immersive environment
- Immersive learning environments for students
- Personalized learning interventions, …

Healthcare
- Planning and Training
- AR-assisted Surgical Interventions
- Pharmaceuticals
- Vein Visualization, …

Real-Estate
- Simulating 3D architectures
- Realistic and Immersive Virtual Tours
- Customized Tours (as per customer needs)
- Adding narrations alongside ambient music, light, & sound effects, …

Gaming & Entertainment
- Immersive Gaming
- Enhanced Entertainment and Social Networking
- Virtual Meetings and Workspaces, …

Fig. 5. Various potential AI-XR metaverse applications.

Computer Vision
- Avatar generation,
- Object detection and recognition,
- Emotion recognition,
- Semantic segmentation,
- Scene understanding,
- Understanding users' activities, …

Natural Language Processing
- Speech-to-text,
- Text-to-speech,
- Chatbots,
- Text/speech-based emotion recognition
- Voice and textual commands, …

Network Communication
- Intrusion detection,
- Intelligent resource allocation,
- Resource slicing,
- Channel state estimation,
- Network and traffic modulation, …

Role of AI in Metaverse

Digital Twin
- Using AI to intelligently monitor and translate data from complex physical systems to digital replica,
- AI-empowered data analytics, and simulations,
- AI-based virtual data generation, …

Blockchain
- Intelligent smart contracts,
- AI-based clustering and classification,
- Knowledge discover and learning,
- Data management,
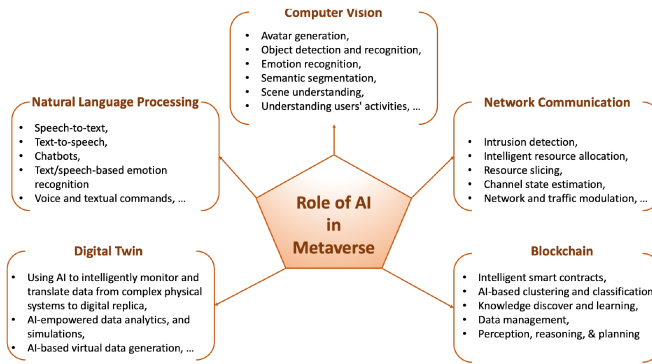- Perception, reasoning, & planning

Fig. 6. Applications of AI in metaverse.

use of face filters in apps such as Snapchat, and games such as the addictive Pokémon GO game, where players could "see" Pokémon characters on the street. Modern mobile phones supporting Lidar technology now can support AR with new software development kits emerging such as Google's AR development platform ARCore, which provides the ability to track motion, understand the environment, and estimate light—three capabilities essential for AR. AR pioneer Louis Rosenberg predicts that within 10 years, most people will be spending more than two hours every day in VR, and augmented reality interfaces will replace mobile phones as our primary interface with digital content.

2.4 Role of AI and ML in Metaverse

Different AI techniques including **machine learning (ML)** and **deep learning (DL)** have many potential applications in the metaverse (as shown in Figure 6). For example, one of the most fascinating features of the metaverse will be voice-based commands that will utilize different voice recognition and analysis techniques for language processing and understanding human commands. In addition, metaverse will use different ML/DL-based regression and classification for data management and decision-making. Also, to provide immersive experiences to the users it will use different generative models to generate photo-realistic avatars and for 3D reconstruction of objects from 2D images. Below, we discuss the potential applications of ML/DL in the metaverse across five dimensions.

2.4.1   Applications of ML in Natural Language Processing (NLP). NLP consists of different techniques that are used for automatically analyzing and understanding human languages (i.e., text and speech). There are many NLP applications that will be part of AI-XR metaverse applications, e.g., speech-to-text, text-to-speech, chatbots, and text/speech-based emotion recognition are the most prominent features of the metaverse. In particular, NLP techniques will be used for the recognition and understanding of complicated human conversations and commands. A key driving force behind the success of NLP methods is the advancement in ML/DL, with the development of new techniques such as **recurrent neural networks (RNN)**, **long-short term memory (LSTM)**, and transformer networks [103].

2.4.2   Applications of ML in Vision. Machine vision or computer vision will be a fundamental component of AI-XR metaverse applications. Different computer vision applications will enable various functionalities in metaverse applications, for example, processing visual data from different sensors to infer high-level visual semantics. The major tasks of the visual processing pipeline include understanding users' activities, emotion recognition, object detection, scene understanding, semantic segmentation, avatar generation, and so on. In addition to these applications, the metaverse is expected to have AI-empowered quality assessment capabilities, e.g., for satisfying the users' demands about viewing high-resolution videos [68]. In this regard, advanced AI methods can be used to develop quantitative and qualitative benchmarks for visual quality assessment.

2.4.3   Applications of ML in Network Communication. Metaverse is expected to simultaneously entertain a massive number of users with the metaverse services provisioned mainly through wireless networks. Over the past few years, substantial research attention has been devoted to improving the overall throughput and performance of wireless network communication, and the use of different AI techniques is the main driving force behind this innovation [32]. Metaverse will mainly include real-time multimedia services that require a reliable connection, high throughput, and low latency to ensure a seamless user experience. Therefore, it is expected that the metaverse will benefit from 5G and beyond empowered communication. The potential of different AI techniques has already been demonstrated for 5G and 6G, e.g., intelligent resource allocation [129], solving resource slicing problem [12], channel state estimation [91], network modulation [140], and so on.

2.4.4   Applications of ML in Blockchain. Service providers in the metaverse will provide users with different incentives in terms of digital assists (e.g., coins) for different events, games, and creative activities. The dispersion of such assets requires a transparent way to record and track such transactions. In this regard, smart contracts–empowered blockchain technology can be leveraged that allows critical information to be stored on an immutable and impenetrable ledger. The decentralized nature of blockchain imbues it with great potential to address security and privacy issues in the metaverse [27]. This potential increases with AI-empowered blockchain applications [167], for example, different AI-based clustering and classification techniques can be used for data analysis stored on blockchain [137]. In addition, different AI techniques can be used for knowledge discovery and learning, efficient data management, perception, reasoning, and planning.

2.4.5   Applications of ML in Digital Twin. The term digital twin refers to the digital replica (i.e., representation) of real objects. A digital twin is capable of synchronizing regular actions, operations processes, and assets with the real world, e.g., analyzing, monitoring, predicting, and visualizing [138]. The digital twin also acts as a bridge where the actual world and digital world interact with each other through different IoT devices [31]. The digital twin will be one of the most important building sectors of the metaverse that allows users to access and use services in the virtual world while exactly depicting the real world in a virtual environment. For example,

surgeons and medical experts can create a digital replica of a patient to study and understand the involved complexities before performing his surgery.

3 CHALLENGES IN IMPLEMENTING SECURE AND TRUSTWORTHY AI-XR METAVERSE APPLICATIONS

Despite the significant potential of different AI-XR metaverse applications, there are various challenges related to security, privacy, and lack of trust that can hinder their wide adoption. A few such challenges include privacy breaches, security invasion, user profiling, unfair AI outcomes, and so on. These challenges may directly or indirectly put the users' safety at risk and can influence social acceptability [81]. Moreover, as discussed above, metaverse is the integration of different modern technologies such as AI, blockchain, and 5G/6G, therefore, it is likely that the inherent issues associated with these technologies get translated into the metaverse. In this section, we describe different challenges that can hinder the secure, safe, robust, and trustworthy employment of AI-XR metaverse applications. Specifically, we characterized and discuss these challenges in two dimensions, i.e., challenges associated with the use of AI techniques including ML/DL-based methods and XR-related challenges in the metaverse. We will start by first discussing AI-related challenges.

3.1 AI-associated Challenges in AI-XR Metaverse

Modern AI techniques that include ML/DL-based models suffer from different vulnerabilities that hinder the smooth, safe, secure, and trustworthy use of these methods in critical applications such as healthcare, autonomous vehicles, and AI-XR metaverse applications. Below, we briefly discuss various such challenges.

3.1.1 Privacy Issues. Ensuring the privacy of the end-users will be a major challenge in AI-XR metaverse applications. As these applications are designed to monitor and collect users' data at an unprecedented fine-grained level, in a bid to create a replica of the digital world [45], there is a greater danger and risk of privacy breaches [150]. For example, to create an immersive virtual scene in the metaverse, data from different sensors will be collected and analyzed using AI models, e.g., facial expressions, brain wave patterns, hand movements, eye movements, biometric, and speech data [150]. This raises obvious concerns regarding the privacy of users and opens a new horizon for digital crimes [45]. Users' sensitive information including daily routine activities, personal logs, and schedules will be stored on a server, which ultimately becomes a critical privacy challenge in a publicly distributed network. Such data include body movements, voice, reflexes, and even more critical data that include subconscious and unconscious responses such as eye movements and physiological signals. Features such as eye tracking are readily accessible using commercially available products such as the HTC Vive Pro Eye and Pico Neo VR headsets even though the XR expert Louis Rosenberg recommends banning such features and data collection in non-health-related applications for ethical reasons. In general, the data is collected through on-device sensors at the user site, processed at their devices or nearby local server, and logged as storage in the cloud. Considering the above-mentioned procedure, some malicious attacks can be encountered, which are classified as (i) data collection, (ii) data storage, (iii) data usage, and (iv) user profiling. Furthermore, as AR technology depends on the precise localization of users in the physical world, modern smartphones use Lidar sensors and **Simultaneous Localization and Mapping (SLAM)** algorithms to precisely locate the user in the real world. This opens up the possibility of privacy violations, as sensitive information may be exploited for anti-social purposes. Some important concerns related to the privacy of data are described next.

Data Collection. Generally, the data in AI-XR metaverse applications is collected through users' input either by voice command or textual input and multi-modal sensors including camera, microphone, textual commands, gesture sensors, and wearable devices. These devices are expected to frequently collect personal information such as daily routine activities, voice and biometric data, and personal preferences including shopping items, TV shows, and preferred food choices. In addition, the private data will be used for the creation of avatars for a digital representation of a real human in the metaverse, which also raises privacy issues. For example, the built-in location sensor in the Oculus headset can be used for tracking users' presence in the real environment with a precise accuracy [150].

Data Storage. The ultimate aim of developing personalized AI-XR metaverse applications is to aid human beings and ease their daily routine activities. These applications will contain multiple sensing devices that generate a substantial amount of data. Whereas these devices will be resource-constrained with small storage units, which leads to the tradeoff between data generation and storage at the edge level. To address these shortcomings, these devices upload their data along with the corresponding logs to online local or global data centers. Though, the data storage tradeoff is resolved by connecting with the servers. However, it also raises privacy concerns regarding the access permissions and data protection of the consumers. Also, if the communication channel is hacked by an attacker, then the data can be manipulated to get the intended outcomes. The literature suggests that adversaries can extract information regarding the actual data even if the communication is encrypted and can track the location of the users by realizing different attacks such as advanced inference attacks [152] and differential attacks [153].

Data Usage and Consent. Continuous data collection through multi-modal sensors including cameras, microphones, and other sensors will be closely involved in the daily routine activities of metaverse users. These sensors collect continuous data regardless of privacy awareness, which is logged over the local or global server(s). Consequently, this procedure raises legal questions regarding the users' consent and the kind of data that is collected and shared. Moreover, metaverse service providers can also utilize this data to optimize their inference models and to make them robust and more personalized. However, it also raises privacy concerns such as the collection, disclosure, and sharing of the data without the explicit or implicit consent of users.

User Profiling. Similar to the current social media (in which users are considered as the product), everything will be a product alike in the metaverse. Metaverse will act as a meta-platform for different entities (such as users, developers, content creators, businesses, etc.), and this raises questions about data collection and its utilization for user profiling [40]. Also, the provisioning of the metaverse services requires that the users should be uniquely identifiable in the metaverse. For this purpose, VR headsets/glasses or any other such device can likely be used for illegally tracking users in real life [127]. Moreover, such devices can be attacked by malicious actors and can be exploited to track users' real locations for possible digital and real-world crimes. Guzman et al. [38] presented a data-centric perspective to avoid unprecedented privacy challenges related to data collection and its usage in MR.

3.1.2 Lack of Trustworthiness. According to the definition of Trust, expressed by Lee and See [80], in the perspective of automated systems, "*Trust is an attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability.*" Metaverse is a data-driven technology in which service providers will use different automated tools to assist human beings as a recommendation engine in various domains, e.g., shopping recommendations, movie recommendations, and even recommendations regarding their health. The efficacy of such recommendation systems is highly dependent on the collection of personalized data for

intelligent decision-making, e.g., an AI model is trained using the collected fine-grained health data to suggest more accurate recommendations regarding the well-being of users. Despite the huge potential of such features of AI-XR metaverse applications, it also raises many questions about which parameters have influenced the underlying AI model in producing a particular decision and the trustworthiness of such predictions and decisions. Next, we discuss such challenges across three dimensions: (i) Truthful AI, (ii) Transparency, and (iii) Explainable and Interpretable AI.

Need for Truthful AI. The evolution in technology also brings threats among them. Over the past few years, AI-based personal assistants including Siri, Alexa, Astro, and the like, have managed to gain wide social acceptance, and these devices are being used by consumers daily for automating household routine tasks. Currently, misinformation, falsehood, or malfunction in AI-based text or speech analyses and command execution are not considered a matter of concern. However, it is expected that AI-enabled intelligent systems with linguistic capabilities will be a major feature of AI-XR metaverse applications. In this regard, it will be quite challenging to enforce the criterion of truthfulness in AI-based systems to ensure the safe selection of statements and behavior according to the social norms of users while interacting with human society.

Transparency Issues. In general practices, developing AI-based systems is about training sophisticated algorithms on large-scale data to learn an efficient and generalizable model that can be deployed in the real-world environment. However, it is a well-known fact that the performance of these models is directly proportional to the transparency of training data, i.e., AI models will perform as well as their training data. Numerous factors such as input data riddled with poor cleansing, the selection of inherently biased data, underfitting, and overfitting influence the performance of these models and can result in fairness and accountability issues (discussed later in this section). Unlike typical application development, there are no quality assurance tools available to spot bugs and evaluate the bias factor in the training data. For instance, if it was known at which stage the model is going to infer at a perfect scale, then there would not be a need to perform training on such large-scale data. This process is all about the hit-and-trial procedure, which is a quite challenging task to identify the right approximations with better data, hyperparameters, and configuration settings.

Explainable and Interpretable AI. The rapid adoption of AI-based applications in human society has also grown the complexity of the systems, which ultimately requires system understandability to make them legitimate and trustworthy. In a critical human-facing technology such as the AI-XR metaverse, interpretable and explainable AI models are required to answer questions about accountability and transparency of their decisions and outcomes: for example, how the employed model reached the decision and which factors influenced the models to make that decision [8]. Such questions are particularly important for human-centric applications where the potential impact of AI will be limited if it is not able to provide accurate and transparent AI predictions. Therefore, the key objective of these models is to develop a relationship of trust between human users and AI. Also, in Reference [88], explainability is described as an inherently human-centric property. However, one of the main challenges in developing explainable methods is the tradeoff between achieving the modesty of an algorithm and ensuring the discretion of sensitive user data. In addition, it is a challenging task to identify the right information while generating a simple yet useful explanation for users. It is worth noting that the terms interpretability and explainability are closely related and are often used interchangeably in machine learning literature, however, these terms are different in practice. Interpretability of the models is defined as the extent to which its outcomes are predictable, i.e., for a given change in the input or model parameter(s), the interpretability enables

us to predict the respective change in its output. On the counter side, explainability deals with the explanation of internal processes of the ML/DL models in a human-understandable way.

In the literature, various studies have been presented that aim to understand and identify the requirements for XAI from a human perspective. For instance, Liao et al. [87] developed an algorithm-informed questionnaire to be asked by various AI developers during semi-structured interviews to identify the gap between the current XAI work and practices to develop user-centered XAI systems. The study concluded that the current research in XAI lacks an understanding of real-world users' needs and requirements for enhancing transparency in AI. Furthermore, the authors suggested a way forward to reduce this gap that includes addressing users' needs and reducing the gap between algorithmic output and developing human-understandable explanations. Similarly, the authors in Reference [88], presented an overview of existing XAI literature from a human-centered perspective. Specifically, they considered the following questions: *What is the role of **human-computer interaction (HCI)** techniques in shaping XAI research?* Furthermore, they highlighted three roles that the HCI technique plays in navigating, assessing, and expanding XAI research that includes: (1) Understanding all AI stakeholders' needs for diverse explainability; (2) Identifying limitations by considering the gap between the explainability and actionable understanding; and (3) Performing theoretical analysis of human explanations, behavioral, and cognitive processes to develop novel computational frameworks to foster enhanced compatibility between humans and XAI. On a similar note, T. Miller emphasized the need to leverage and incorporate insights from various social sciences including psychology, philosophy, and cognitive science to define explanations for XAI [95]. Considering these fields, the author also provided a detailed overview of existing literature related to human explanation along with providing insights about how these existing methods can be incorporated into XAI systems.

3.1.3   ML Security Issues. Despite the state-of-the-art performance of modern AI techniques including DL-based systems, it has been shown that these models are highly vulnerable to carefully crafted adversarial perturbations known as adversarial ML attacks [136]. The threat of these attacks has been already demonstrated for many critical applications such as healthcare, autonomous vehicles, and so on. On a similar note, AI-XR metaverse applications are essentially critical, as they involve humans, and ensuring their safety from any harm is profoundly important. On the counter side, the existence of these challenges raises many concerns about the safety, security, and robustness of AI-based metaverse applications, thus hindering their practical deployment. As it is equally important that any AI-based system should be equally trusted by all stakeholders involved including service providers, developers, and end users. These challenges are detailed later (Section 4).

3.1.4   Lack of Fairness and Accountability. Modern AI methods, like advanced DL models, lack fairness and accountability in their decisions [111]. However, such questions are particularly important for critical applications like AI-XR, in which the model's decisions can have life-threatening consequences for the end-users. Moreover, AI models are developed using training data, which will be mainly collected from human users in AI-XR metaverse applications for providing immersive experiences. Humans possess certain biases that will be readily reflected in the data they generate, and when this data is used for training AI models, the data bias will be directly translated into the developed AI-based system. As a result, the model will be biased towards certain samples that contain certain features (bias), and its decision will not be fair. On a similar note, the critical nature of AI-XR metaverse applications demands accountable decisions. Consequently, data bias if remained unaddressed can ultimately lead to unintended consequences [79].

3.1.5   Identity Theft and Authentication Attacks. Users'/avatars' identities in the metaverse can be stolen or impersonated illegally, leading to authentication and access control issues in the

interconnected virtual worlds. Identity theft in the metaverse will be more dangerous than traditional attacks. The identity of a user once stolen will reveal everything about that person's digital assets, avatars, and social relationships. The attackers can exploit different vulnerable VR gadgets and other service authentication loopholes to realize identity theft attacks and can steal the victim's secret keys of digital assets and bank details. It has been reported that about 17 users in the OpenSea NFT marketplace were hacked through a phishing attack and flaws in the smart contract that resulted in a loss of $1.7 million.[1]

Metaverse will leverage different biometrics and password-driven technologies to authenticate users and their avatars in the virtual worlds. The attacker can evade such authentication systems to impersonate real users' identities to get control of the whole virtual world. Evading AI-based biometric systems has become easier with the advancements in adversarial ML research. Therefore, AI-empowered speech-and-face-recognition-based biometric systems can be easily attacked to realize impersonation attacks. Once the attacker has access to the metaverse, it can exploit the data generated by the victim's devices to deceive him, committing a crime in the virtual space. However, the exposure of biological data when used for authentication purposes can also lead to severe consequences [77]. Moreover, the authentication of social friends of a user using their avatars is much more challenging in the metaverse as compared to real-world identity authentication. In this regard, facial data, voice, and videos can be used to develop an AI-based avatar authentication system, however, the unsolved inherent issues of AI can still hinder its practicality.

3.1.6 The Bias Problem. Bias refers to a model making certain unethical assumptions about the data. Human bias along with its many aspects has been studied by researchers in many disciplines, including law, psychology, and so forth. In Reference [99], bias is defined as "the prejudice or inclination of a decision made by an AI system which is in a way considered to be unfair for or against one person or group." Bias in recommendation systems, advertising algorithms, facial recognition systems, and risk assessment tools has been widely studied in recent years. In metaverse applications, data will be collected from a heterogeneous group of people and sources having their own characteristics, stereotypes, and behaviors, which introduces different biases in the collected data. In References [93, 102, 135], the authors discuss different kinds of bias based on the sources and the types of bias. On the base of sources, these biases have been divided into further categories: biases caused by data, biases caused by algorithms, and biases caused by user interaction.

3.2 XR-related Challenges in Metaverse

AI-XR metaverse applications are essentially human-centric, and ensuring the security, privacy, security, and robustness of such applications is of utmost importance. It has been envisioned that an entirely new form of digital media will emerge from the use of VR and AR in the modern metaverse (TV, print media, and the web). In recent years, there has been immense discussion regarding the concerns about surveillance capitalism, which is happening on the Internet in different applications. Many large tech organizations providing Internet services, such as Facebook, Google, Microsoft, and Amazon, collect large amounts of data related to the surveillance of their users, which is then used to satisfy the needs of advertisers [146]. The pioneers of VR and AR such as Jaron Lanier [78] and Louis Rosenberg[2] have predicted that the concerns about surveillance are expected to rise in metaverse. For example, it has been shown how reconfiguring AR in Pokémon Go (an AR mobile game) drew unexpected audiences to museums and public spaces like trains to fill in the space thus creating a form of virtual trespassing. It has been reported that people were

---

[1]https://threatpost.com/nft-investors-lose-1-7m-in-opensea-phishing-attack/178558/
[2]https://bigthink.com/the-future/metaverse-augmented-reality-danger/

putting their lives in danger to pursue virtual characters. This highlights that safety concerns may arise when such immersive technologies are engineered for gaming and experiences. Below, we discuss the key challenges that are hindering metaverse applications in general and we will later discuss the specific challenges that arise with the use of different AI techniques in metaverse applications (Section 4).

3.2.1 Safety Issues. There are various concerns regarding the mental and physical safety of metaverse users. There are several reported incidents of digital harassment, theft, and bullying in XR applications [92]. The report on "Immersive and Addictive Technologies" highlights rampant incidents of sexual harassment, cyberbullying, and grooming online [100]. Ensuring the safety of users is a major challenge for AI-XR metaverse applications because of the fact that such incidents have real damage and harm to users despite being experienced in the virtual world. The avatars generated using recent advancements in AI techniques, in particular, generative models, can appear more realistic in AI-XR metaverse applications and can engage users in promotional conversation, thus providing a false sense of a real human behind the avatar. The avatars in such a promotional content are fueled with more personalized data (such as your vitals, emotions, expressions, etc.) to look more realistic. Also, these sales avatars can pitch products to you more persuasively than any real salesman or even a recommendation system due to their access to rich cyber-physical data about you. The research in deep fake technology and photorealistic avatars is already at the stage where computer-generated content is indistinguishable from the real. Such advancements can be leveraged to realize an adversarial attack on AI-XR metaverse applications to get the intended behavior and outcomes.

3.2.2 Potential Antisocial Aspects. There are various opinions regarding the antisocial aspects of AI-XR metaverse applications. Many people think that introducing AI-XR metaverse may detract the users (humans) from their purposes and may have a somatic effect. In the literature, it has been shown that extended times online can result in users demonstrating post-VR sadness and detachment from reality. For instance, Aldous Huxley in 1932 wrote in his social science dystopian fiction novel that using technology can lead to self-inflicted harm that can influence people to be diverted from their higher priorities and become more prone to being influenced by other interests. As a result of such a quest for technological utopia, the human psyche and society as a whole are greatly afflicted. Social critics have long argued that various digital media, such as television, the Internet, and the Web, make people docile and less connected to the real world. For example, Jerry Mander in 1978 wrote in his book *Four Arguments for the Elimination of Television* that TV removes the sense of reality from people, promotes capitalism, can be used as a scapegoat, and all these three factors work together negatively. The modern technological disruptions including the web and social networking services have created a filter bubble detaching people from the real world and the truth. Due to these reasons, the current era is also referred to as a post-truth era [51]. We may reasonably expect that alienation from the real world will exacerbate with the increasing adoption of VR, AR, and AI-XR metaverse applications, which aim at changing the human perspective of the world. This argument can be supported by the fact that, in 2018, the World Health Organization formally included "gaming disorder" in its International Classification of Diseases following research that shows that technology can promote addictive behavior in people. Moreover, the literature focused on analyzing the social implications of metaverse argues to understand and identify potential psychological problems that can arise in metaverse [23].

3.2.3 Ethical Aspects. Any technological intervention involving humans suffers from some serious ethical issues, especially the one that contains intelligence. The **Institute of Electrical and Electronics Engineers (IEEE)** has recently published a report on Ethically Aligned Design that

mainly focuses on the Ethics of Autonomous and Intelligent Systems [126]. This report emphasizes the need of developing ethical **autonomous and intelligent systems (A/IS)** that promote human well-being and protect human rights through transparent and accountable A/IS and the prevention of the misuse of AI. This report is the collective effort of hundreds of researchers having diverse backgrounds and expertise in important areas such as governance, technology, civil society, and policy-making. This report has a dedicated section on XR and, interestingly, IEEE also has a Global Initiative on Ethics of XR.[3] In this report, various ethical issues related to XR have been highlighted, including users' preference for virtual life over the real world and complete disengagement with society. In addition, the reports conclude with the following remark regarding XR: *"The nature of XR environments fosters unique legal and ethical challenges that can directly affect users' privacy, identity, and rights. Society will need to rethink notions of privacy, accessibility, and governance across public and private spaces. New laws or regulations regarding data ownership, free use, universal access, and adaptive accessibility within XR environments may need to be developed."*

3.2.4 Regulatory Challenges. The big tech companies have resisted regulation, decrying the fact that regulation will slow down innovation. However, there are various ethics researchers and social scientists who are arguing for much greater regulation to ensure that consumer rights are protected. In this regard, the EU **General Data Protection Regulation (GDPR)** has paved the way for many countries and regions attempting to develop similar regulatory laws to protect Internet users. These regulations mainly emphasize the importance of the non-profiling of users (i.e., limiting the storage of tracking data) and for better transparency (i.e., online services and applications should specify why and what information is being stored). On a similar note, AI-XR metaverse applications are subject to the requirement of being transparent in terms of data collection and utilization and should also be subject to the informed consent of users. Also, the development of such applications requires thoughtful deliberation from a regulatory perspective. For instance, it is worth considering banning non-medical applications to collect vital biomedical statistics due to the high risks of being exploited maliciously. To mitigate the risk of users being manipulated deceptively, metaverse operators may be bounded to transparently declare the staging of virtual products and experiences in the metaverse. Rosenberg, one of the pioneers of VR and AR, has already started to argue about the need for regulation for metaverse applications [118]. For instance, he suggested leveraging the arguments regarding the regulation of social media for developing a legal and philosophical basis for metaverse regulation. As the metaverse can be deemed as an evolutionary expansion of similar services. Rosenberg argued that the only solution to eliminate ethical and privacy-related concerns associated with metaverse is to shift from an advertising-based to a subscription-based business model in which users pay a subscription fee for accessing the metaverse platform. This eliminates the service providers' need to monitor their user base to a greater extent, however, this is not a feasible solution, as it is difficult to say whether or not users will pay for a safer metaverse.

4 ANALYZING SECURITY AND TRUSTWORTHINESS ASPECTS OF AI-XR

In this section, we will discuss the challenges associated with the use of different AI techniques (in particular, ML/DL-based models) that hinder the safe, secure, and trustworthy deployment of these methods in metaverse applications. We start by first providing a broad overview of AI security in the metaverse.
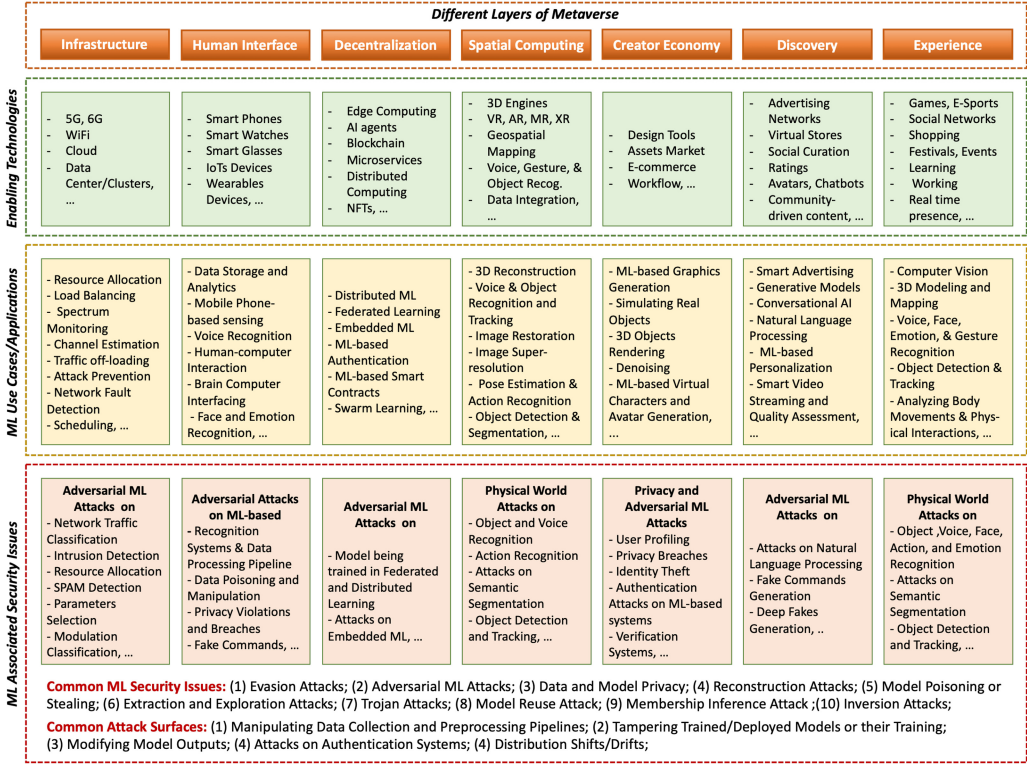
---

[3]https://standards.ieee.org/industry-connections/ethics-extended-reality/

Different Layers of Metaverse

	Infrastructure	Human Interface	Decentralization	Spatial Computing	Creator Economy	Discovery	Experience
Enabling Technologies	- 5G, 6G - WiFi - Cloud - Data Center/Clusters, ...	- Smart Phones - Smart Watches - Smart Glasses - IoTs Devices - Wearables Devices, ...	- Edge Computing - AI agents - Blockchain - Microservices - Distributed Computing - NFTs, ...	- 3D Engines - VR, AR, MR, XR - Geospatial Mapping - Voice, Gesture, & Object Recog. - Data Integration, ...	- Design Tools - Assets Market - E-commerce - Workflow, ...	- Advertising Networks - Virtual Stores - Social Curation - Ratings - Avatars, Chatbots - Community-driven content, ...	- Games, E-Sports - Social Networks - Shopping - Festivals, Events - Learning - Working - Real time presence, ...
ML Use Cases/Applications	- Resource Allocation - Load Balancing - Spectrum Monitoring - Channel Estimation - Traffic off-loading - Attack Prevention - Network Fault Detection - Scheduling, ...	- Data Storage and Analytics - Mobile Phone-based sensing - Voice Recognition - Human-computer Interaction - Brain Computer Interfacing - Face and Emotion Recognition, ...	- Distributed ML - Federated Learning - Embedded ML - ML-based Authentication - ML-based Smart Contracts - Swarm Learning, ...	- 3D Reconstruction - Voice & Object Recognition and Tracking - Image Restoration - Image Super-resolution - Pose Estimation & Action Recognition - Object Detection & Segmentation, ...	- ML-based Graphics Generation - Simulating Real Objects - 3D Objects Rendering - Denoising - ML-based Virtual Characters and Avatar Generation, ...	- Smart Advertising - Generative Models - Conversational AI - Natural Language Processing - ML-based Personalization - Smart Video Streaming and Quality Assessment, ...	- Computer Vision - 3D Modeling and Mapping - Voice, Face, Emotion, & Gesture Recognition - Object Detection & Tracking - Analyzing Body Movements & Physical Interactions, ...
ML Associated Security Issues	**Adversarial ML Attacks on** - Network Traffic Classification - Intrusion Detection - Resource Allocation - SPAM Detection - Parameters Selection - Modulation Classification, ...	**Adversarial Attacks on ML-based** - Recognition Systems & Data Processing Pipeline - Data Poisoning and Manipulation - Privacy Violations and Breaches - Fake Commands, ...	**Adversarial ML Attacks on** - Model being trained in Federated and Distributed Learning - Attacks on Embedded ML, ...	**Physical World Attacks on** - Object and Voice Recognition - Action Recognition - Attacks on Semantic Segmentation - Object Detection and Tracking, ...	**Privacy and Adversarial ML Attacks** - User Profiling - Privacy Breaches - Identity Theft - Authentication Attacks on ML-based systems - Verification Systems, ...	**Adversarial ML Attacks on** - Attacks on Natural Language Processing - Fake Commands Generation - Deep Fakes Generation, ...	**Physical World Attacks on** - Object ,Voice, Face, Action, and Emotion Recognition - Attacks on Semantic Segmentation - Object Detection and Tracking, ...

Common ML Security Issues: (1) Evasion Attacks; (2) Adversarial ML Attacks; (3) Data and Model Privacy; (4) Reconstruction Attacks; (5) Model Poisoning or Stealing; (6) Extraction and Exploration Attacks; (7) Trojan Attacks; (8) Model Reuse Attack; (9) Membership Inference Attack ;(10) Inversion Attacks;

Common Attack Surfaces: (1) Manipulating Data Collection and Preprocessing Pipelines; (2) Tampering Trained/Deployed Models or their Training; (3) Modifying Model Outputs; (4) Attacks on Authentication Systems; (4) Distribution Shifts/Drifts;

Fig. 7. Overview of ML security in metaverse.

4.1 Security of ML in Metaverse

The impact of metaverse applications will be social and economic, and these applications will be more susceptible to undesirable adversarial action(s). AI will be the fundamental driving force behind the success of the metaverse, and there are numerous applications of AI in different layers of the metaverse. However, the use of AI algorithms in AI-XR metaverse applications also opens them up to different adversarial attacks. In Figure 7, we highlight the threat of different security and privacy attacks that can be realized in different applications in almost every layer of the metaverse. The figure also highlights that there are various common ML security issues and attack surfaces that get shared across the architectural landscape of the metaverse across different AI applications at each level. In this section, we discuss different AI-associated security and privacy attacks on AI-XR metaverse applications.

4.2 Potential Attacks on AI-based Metaverse Applications

The threat of adversarial ML attacks has already been shown to be successful in compromising the integrity of AI techniques in many critical tasks, e.g., connected and autonomous vehicles [112], computer vision [4], and healthcare [111], just to name a few. Furthermore, AI-XR algorithms could be biased either due to data imbalance or adversarial subversion. Many of the ethical dilemmas and social harms such as distraction, narcissism, disinformation, outrage, and polarization stem from the economic model of surveillance capitalism in which service providers give the customers everything and anything that makes the company money. In this way, these companies pander

to the base animal desires of people and exploit their cognitive biases, effectively downgrading humans and manipulating them for ulterior selfish purposes.

In the adversarial ML literature, an adversarial example is defined as the input to the deployed AI model crafted by an adversary by introducing imperceptible noise into the legitimate sample to get the intended outcomes. In general, there are two types of adversarial ML attacks: (1) poisoning attacks that aim at altering the training process of the AI model; and (2) evasion attacks that are focused on evading the deployed AI model by making inferences (they are also known as inference time attacks). In poisoning attacks, the adversary mainly modifies the training data to tamper with the learning of the AI model [17]. In contrast, test data is manipulated in evasion attacks to get the desired predictions from the model [16]. Recent works have shown that AI models are vulnerable to attacks at both training and inference stages [75]. Training stage attacks typically corrupt a small subset (typically ~1%) of the training data samples to achieve malicious goals during AI model training [11, 57]. However, the inference stage attacks cause a trained model to misbehave on adversarially crafted test inputs [9, 10].

Attacks on AI models are generally carried out by first defining a threat model. A threat model is a set of assumptions regarding the attackers' abilities to access and affect a typical AI model training pipeline. Broadly, there are two main threat models—the poisoning threat model (i.e., realizing poisoning attacks) and the adversarial threat model (realizing evasion attacks). A poisoning threat model assumes an attacker who can control a small set of the training dataset to adversely affect the training of the model. An adversarial threat model assumes an attacker who can access and, to a certain extent, perturb the inputs to an already trained AI model. In the following, we highlight major security threats associated with the use of AI techniques in metaverse applications that include computer vision, NLP, network communication, authentication, and recognition systems.

4.2.1 ML-associated Security Issues in Computer Vision. Computer vision is one of the central building blocks in the foundation of the metaverse. In recent years, DL algorithms have enabled major advances in computer vision ranging from image classification [25, 39] to scene understanding [26, 36] and generating realistic images [158]. However, the discovery of the adversarial vulnerabilities of DL-based image processing models by Szegedy et al. [136] sparked a growing concern regarding the reliability and security of these deep models [7, 76, 143, 145]. Numerous works have analyzed these adversarial vulnerabilities in greater depth under different threat models [75]. In general, adversarial attacks work by optimizing the perturbation, Δx, to an input image, x, such that the output of the model, $\mathcal{F}$, is significantly changed, maximize $||\mathcal{F}(x) - \mathcal{F}(x + \Delta x)||$. Δx is typically optimized based on the gradients, which are either computed directly (white-box scenarios) or estimated by introducing random noise (black-box scenarios). A summary of various adversarial ML attacks on different computer vision applications can be seen in Table 2.

4.2.2 ML-associated Security Issues in NLP. Similar to the vulnerabilities of ML models for vision applications, the literature demonstrates that the ML methods for modeling NLP tasks are also vulnerable to malicious attacks, at both the training and the inference stages of a typical ML pipeline [9, 11]. Below, we discuss such attacks.

Poisoning and Trojaning Attacks: Poisoning attacks and trojaning (also known as backdoor) attacks are the most widely known training stage attacks in NLP. Poisoning attacks aim to tamper with the training of the model so it is unable to perform satisfactorily on the test inputs [17, 96, 134]. Trojaning attacks aim to insert a trojan—typically characterized by a specific pattern of words known only to the attacker in the input sequence—into a model such that the model behaves normally on natural test inputs but malfunctions as desired by the attacker (through the use of the trojan pattern of words [57, 125]). Similar to the case with the computer vision applications, the

Table 2. Summary of Different Adversarial Attacks on Various Computer Vision Applications (that Are Expected to be Potential Metaverse Applications)

Application	Authors	Methodology	Datasets	Before → After
Face Authentication	Goswami et al. [55]	Studied how different architectures affect adversarial vulnerabilities.	MEDS, PaSC	89.3% → 41.6%
	Sharif et al. [128]	Developed adversarial glasses to fool face recognition systems.	Celebrity Face	98.95% → 0%
	Shen et al. [130]	Developed black-box attack for face recognition systems using visible light.	CusFace, LFW	100% → 7.9%
	Chatzikyriakidis et al. [30]	Perturbed facial images to fool automatic face recognition to secure a person's identity.	CelebA	97.8% → 4%
	Dabouei et al. [37]	Studied the vulnerability of face recognition systems against geometrically perturbed faces.	VGGFace2	100% → 0.14%
	Zhong et al. [177]	Used dropout and feature-level attacks to improve the transferability of adversarial inputs.	VGGFace2	100% → 3.24%
	Dong et al. [41]	Used evolutionary algorithm to find adversarial inputs against the models' decisions.	LFW	100% → 0%
	Wenger et al. [155]	Proposed improved physically realizable attack against face recognition.	VGGFace	100% → 10%
	Ali et al. [11]	Proposed multi-trigger backdoor attack against backdoor defenses.	Celebrity Face	88% → 8%
	Xue et al. [166]	Exploited hidden facial features as triggers of the backdoor attack.	VGGFace	100% → 0.02%
Object Detection	Zhang et al. [173]	Proposed generalizable contextual adversarial perturbations against object detectors.	PascalVOC, COCO	78.8% → 1.6%
	Lee et al. [82]	Showed that non-overlapping physical patches can fool object detectors.	COCO	55.4% → 0.05%
	Xie et al. [161]	Proposed multi-targeted adversarial attacks to fool object detectors.	PascalVOC	72.07% → 3.36%
	Xie et al. [161]	Showed that multi-targeted adversarial attacks against object detectors are transferable.	PascalVOC	54.87% → 37.9%
	Wei et al. [154]	Utilized generative methods to efficiently obtain transferable adversarial inputs.	PascalVOC	43% → 3%
	Wang et al. [151]	Utilized position and label information to attack black-box object detectors.	PascalVOC	100% → 16%
	Wu et al. [157]	Leveraged natural rotations to insert a backdoor into the object detectors.	PascalVOC	89.5% → 4.45%
3D-Object Modelling	Wang et al. [148]	Optimally generates adversarial perturbations against 3D-Object detectors.	KITTI	84% → 0%
	Xiang et al. [159]	Generated 3D adversarial point clouds against Point-Net model.	ModelNet40	93% → 0%
	Hamdi et al. [61]	Exploited an auto-encoder to generate transferable 3D adversarial perturbations to point cloud.	ModelNet40	93% → 5%
	Meloni et al. [94]	Used off-the-shelf 3D surrogates to transfer attack on 3D object models.	N/A	100% → 0%
	Li et al. [85]	Proposed a novel formulation to develop backdoor triggers against 3D point cloud models.	ShapeNetPart	98.4% → 0.5%
Semantic Segmentation	Arnab et al. [13]	Performed an in-depth study of adversarial vulnerabilities of semantic segmentation models.	Cityscapes	77.1% → 19.3%
	Xie et al. [161]	Proposed multi-targeted adversarial attacks to fool semantic segmentation models.	PascalVOC	72.07% → 3.36%
	Hendrik et al. [64]	Analyzed universal adversarial perturbation to fool a segmentation model for any input.	Citscapes	64.8% → 12.9%
	Li et al. [86]	Poisoned the segmentation models using object-level target class and semantic triggers.	ADE20K	37.7% → 25.2%
	Feng et al. [48]	Proposed frequency-injection backdoor attack against medical image segmentation tasks.	KiTS-19	54.5% → 21.1%

trojaning attacks, being more difficult to be detected as compared to the poisoning attacks, pose a greater threat to NLP applications in the metaverse.

Adversarial Attacks. Although adversarial examples have been extensively studied in computer vision, they have received significantly limited attention in NLP tasks mainly due to the discrete input search space—minimal adversarial perturbations in the input are no longer feasible in NLP [8, 9]. Recently, however, there have been numerous works highlighting the adversarial vulnerabilities of the NLP-based ML models. Notable adversarial attacks include Text-bugger, Text-fooler, PWWS, and **BERT Adversarial Example (BAE)**.

Adversarial attacks against NLP models generally follow three major steps—evaluation, perturbation, and selection—to achieve some adversarial goal—for example, targeted or untargeted misclassification—under a predefined threat model. Consider, for example, an input sequence $X = \{x_1, x_2, \dots, x_i, \dots, x_n\}$ correctly classified by an NLP model, $\mathcal{F}$, in class, $\mathcal{F}(X) = y \in \mathbb{R}^M$. At

Table 3. Summary of Different Adversarial ML Attacks on Various NLP Applications

Application	Authors	Methodology	Datasets	Before → After
Language to Language Modelling	Zhand et al. [175]	Proposes word saliency speedup local search method to attack translation machines.	NIST (MT)	92.48% Degradation
	Boucher et al. [22]	Uses invisible characters, homoglyphs, and deletion control characters to fool the model.	WMT14	37% → 1%
Fake-news Detection	Li et al. [84]	Exploits BERT-MLM to fool a fine-tuned BERT model by generating coherent perturbations.	AG news	94.2% → 10.6%
	Ali et al. [10]	Proposes an adaptive adversarial attack to generate perturbations against statistical defenses.	Kaggle Fake news	95% → 0%
	Jin et al. [71]	Identifies key contributing words and replaces them with synonyms while retaining the coherence.	Kaggle Fake news	96.7% → 15.9%
	Zellers et al. [172]	Presents a generative model, Grover, to generate fake news that fools fake-news detectors.	Not applicable	95% → 67%
	Pan et al. [104]	Exploits linguistic styles as triggers to backdoor an NLP model.	COVID	95.1% → 6.7%
Toxicity/Sentiment Classification	Garg et al. [53]	Exploits BERT-MLM to generate adversarial perturbations that are coherent with the context.	Amazon	96% → 11%
	Boucher et al. [22]	Uses invisible characters, homoglyphs, and deletion control characters to fool the model.	Wikipedia Detox	95% → 19.5%
	Ebrahimi et al. [44]	Leverages atomic flip operation to swap tokens to fool NLP fake news classifiers.	AG news	92.35% → 27.7%
	Li et al. [83]	Exploits model gradients to find and perturb the most positively contributing words.	IMDB	90.7% → 0%
	Li et al. [84]	Exploits BERT-MLM to fool a fine-tuned BERT model by generating coherent perturbations.	IMDB	90.9% → 11.4%
	Jin et al. [71]	Identifies key contributing words and replaces them with synonyms while retaining the coherence.	Yelp	93.8% → 1.1%
	Chen et al. [33]	Uses char- and word-level triggers to backdoor NLP sentiment classifiers.	SST-5	55% → 0%
	Irtiza et al. [70]	Proposes a context-aware hidden trigger backdoor attack against NLP classifiers.	IMDB	84.5% → 2.48%

the evaluation stage, the attacker uses some *impact scoring function* to compute a set, I_x, representing the impact of each word over the output. At the perturbation stage, the attacker repeatedly perturbs the most impactful words in I_x using some pre-defined *perturbation mechanism* such that the semantic and contextual value of X remains preserved. At the selection stage, the most optimal perturbation is selected. Table 3 provides a summary of various adversarial ML attacks on different NLP applications.

4.2.3 ML-associated Security Issues in Networking. AI-XR metaverse applications will provide ubiquitous connectivity to a massive number of users over wireless networks. Over the past few years, many AI-based algorithms have been developed to improve the performance of wireless communication and networking systems that will be used in different layers of network architecture [32]. The use of AI in wireless communication empowers wireless devices to perform many important intelligent functions such as network composition, analyzing traffic patterns, managing content requests, analyzing wireless channel dynamics, and so on. Moreover, AI-based algorithms have been used for optimizing different network constraints such as high throughput and low latency for different multimedia applications. A prominent use case is to leverage intelligent proactive load management in 5G and 6G communication networks and predictive data analytics to improve network operations. Despite the significant potential of using various AI algorithms for different optimizing applications in wireless networks, recent studies have highlighted that AI models are highly susceptible to adversarial ML attacks. For instance, Usama et al. [141] used a **generative adversarial network (GAN)** for realizing adversarial attacks on network intrusion detection. The threat of adversarial ML attacks on network traffic classification is demonstrated in Reference [144] and for cognitive self-driving networks is presented in References [143, 145]. Similarly, the threat of adversarial ML for 5G networks is analyzed in Reference [142]. A summary of various adversarial ML attacks on network applications is presented in Table 4. We refer interested readers to a detailed survey highlighting the threat of adversarial ML in network security [69].

In addition to the above-motioned adversarial vulnerabilities associated with the use of AI techniques in many network applications, some other critical network-related issues can hinder the

Table 4. Summary of Different Adversarial ML Attacks on Different Networking Applications

Application	Authors	Methodology	Datasets	Before → After
Intrusion Detection	Usama et al. [141]	Exploited GAN to craft adversarial examples to evade intrusion detection model.	KDD99	89.12% → 56.55%
	Aiken et al. [3]	Perturbed a few features to evade four ML classifiers trained for detecting DDoS attacks.	KDD99	100% → 0%
Network Traffic Classification	Usama et al. [144]	Crafted adversarial examples using mutual information in black-box settings.	UNB-CIC Tor Data	96% → 77%
Modulation Classification	Usama et al. [142]	Used C&W attack to evade traffic modulation classifier.	RML2016.10a	85% → 15%
	Sadeghi et al. [121]	Realized white-box and black-box attacks on VT-CNN model using PCA-based perturbations.	GNU Radio	75% → 38%
Network Modulation	Usama et al. [142]	Realized black-box attack on channel autoencoder on unsupervised and DRL models.	RML2016.10a	95% → 80%
Malware Classification	Usama et al. [143]	Realized three SOTA adversarial ML attacks, i.e., FGSM, BIM, and JSMA.	Malware Image Data [145]	98.39% → 1.87%
Abnormal KPI Detection	Usama et al. [145]	Leveraged two SOTA attacks to evade ML-based abnormal KPI detection classifiers.	LTE network data	98.8% → 13.7%
Channel State Estimation	Sagduyu et al. [122]	Realized three attacks: spectrum poisoning, jamming, and priority violation.	Not Articulated	95.58% → 23.12%

smooth operation of the metaverse at a global level. For instance, centralized network architecture provides flexibility in terms of cost-saving, simplicity, and ease in performing different operations. However, such architectures are more prone to a **single point of failure (SPoF)** and **distributed denial of service (DDoS)** attacks [149]. For example, if a powerful attacker gets control of the network, then it may lead to severe challenges such as SPoF and DDoS. To address such issues, the literature suggests leveraging decentralized network architecture [97]. In addition, decentralization will potentially amplify the transparency and trust of users in exchanging their virtual belongings (such as digital assets and virtual currencies) among each other and across different virtual worlds in the metaverse. However, many issues arise with the use of decentralized approaches, e.g., reaching a consensus on an ambiguous operation among the huge number of entities in a dynamic metaverse. *DDoS:* Metaverse will include a massive number of IoT devices, which can be compromised by an attacker to form a botnet to realize DDoS attacks [15]. *Sybil Attacks:* In a Sybil attack, the adversary pretends to have fake (or manipulated) identities of legitimate users or devices. Using such stolen identities, he can take over the network.

4.2.4 Security Issues in Cloud-hosted ML Models. Outsourcing the training of ML/DL models to third-party services that offer powerful computational resources on the cloud is prevalent nowadays. These services allow ML developers to upload their data and models for training over their cloud platforms. It is expected that such services will be featured in AI-XR metaverse applications, as they provide the flexibility of developing AI models using sufficiently large training datasets while reducing the cost and time. However, the literature demonstrates that such services are vulnerable to a variety of attacks, such as backdoor attacks [34], exploration attacks [124], model inversion [168] and model extraction attacks [74], and so on. More details about various attacks and defenses for cloud-hosted ML models can be found in Reference [109]. Visual illustration of adversarial ML attacks on different potential applications in the AI-XR metaverse is presented in Figure 8.

4.3 Attacks on VR

The literature highlights that VR systems are vulnerable to adversarial attacks. For instance, Casey et al. [29] demonstrated that humans in VR systems can be controlled like joysticks—thus providing
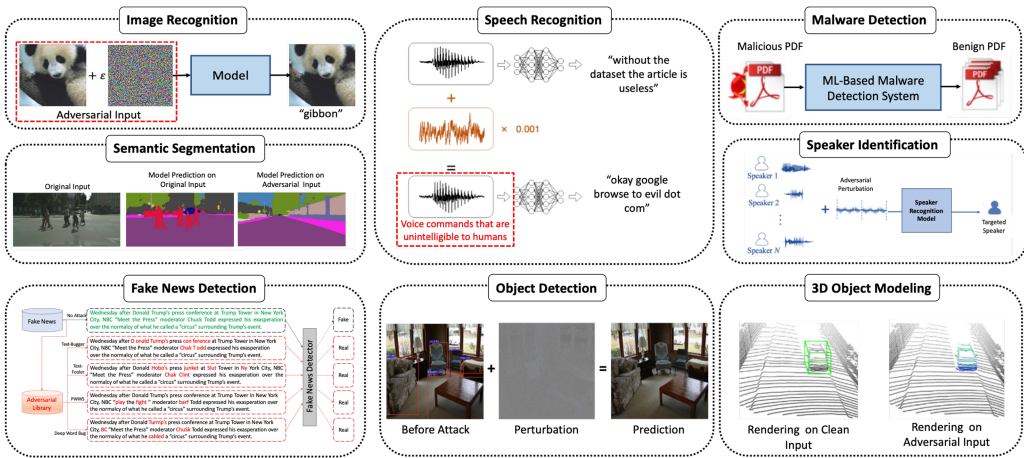
Fig. 8. Illustration of adversarial ML attacks on different potential applications in AI-XR metaverse. Individual figure references: Image Recognition [136]; Speech Recognition [113]; Malware Detection [165]; Semantic Segmentation [50]; Speaker Identification [162]; Fake News Detection [9]; Object Detection [151]; and 3D Object Modeling [148].

the adversary the ability to control the movement of VR user without his consent or getting into his knowledge. Moreover, the literature suggests that both security and privacy attacks can be realized on VR/XR systems [147]. Therefore, developing secure and robust AI-XR metaverse systems is crucial to the widespread adoption of metaverse applications that are not vulnerable to adversarial attacks or are capable to withstand such attacks and mitigate their impact.

4.4 Analyzing Implications of ML Security, Privacy, and Trust Issues: An AI-XR Case Study

In this section, we present an ML/DL-based pipeline for a potential AI-XR metaverse application use case. We then analyzed various challenges and threats that can arise at each development stage. The pipeline is developed while considering a general metaverse application—a virtual conference, in which the participants are remotely connected from different places (the pipeline is presented in Figure 9). A unique avatar is representing each participant while each avatar is expected to reflect real-time voice, facial expressions, and gestures. The voice of each participant is translated into the native language of all the participants along with generating the transcription. The pipeline depicts different ML/DL-empowered tasks: (1) *3D/4D Visual Reconstruction*—responsible for generating photo-realistic avatars; (2) *3D Visual Mapping*—to reflect real-time multi-modal expressions (i.e., audio, facial, and gestures, etc.); (3) *Speech Recognition and Synthesis*—to interpret and translate the voice of recipient into other languages; and (4) *Speech-to-text Synthesis*—to generate the transcription of audio conversations of all members.

As depicted in Figure 9, data acquisition is performed by collecting raw audio input through a microphone for NLP, whereas, depth cameras and laser scanners are used for 3D/4D visual reconstruction and mapping. In the next step, acquired data is pre-processed through several techniques, including data denoising, deblurring, silence removal, and so on. The processed data is then labeled for the training of ML/DL models in a supervised/semi-supervised learning fashion. After successful data preparation, 3D visual mesh construction and segmentation models are trained to perform 3D avatar reconstruction. However, acoustic and language models along with neural vocoders are trained to perform multilingual translation and transcription tasks. Although, in the
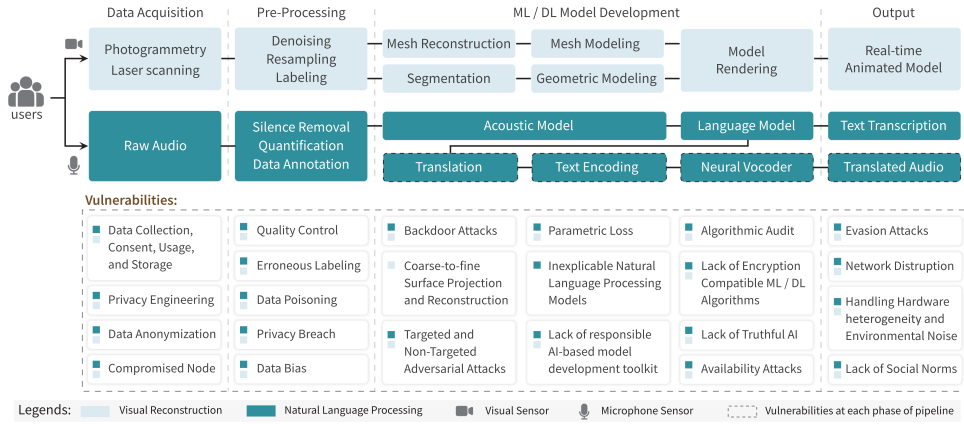
Fig. 9. A prospective **pipeline for developing AI-XR metaverse applications** for multi-lingual communications, involving various security challenges at each stage.
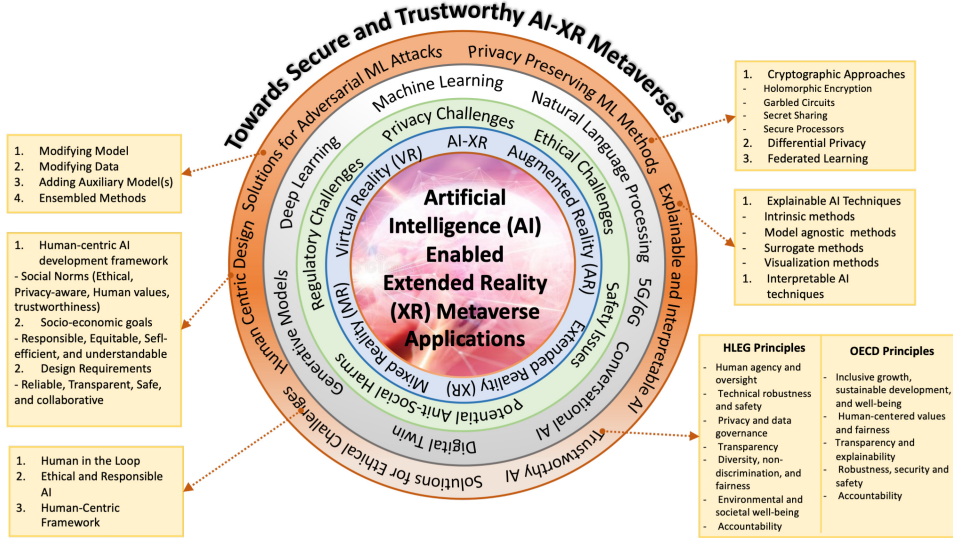

Fig. 10. An abstraction of different ML-associated challenges along with a taxonomy of various solutions that can be used to address those challenges.

literature, these pipelines have demonstrated significant performance in 3D reconstruction and NLP tasks, however, they are highly exposed to various privacy and security attacks at each stage of the development pipeline (as shown in Figure 9).

5   TOWARDS DEVELOPING SECURE AND TRUSTWORTHY AI-XR

The development of secure, safe, and trustworthy AI-XR metaverse applications is fundamentally very important, in this section, we will discuss different potential solutions that can be leveraged to address challenges associated with the use of AI in particular and for the overall system in general. An abstraction of different techniques that can be leveraged to address the ML-associated issues is shown in Figure 10, and these methods are described next.

5.1 Solutions for Privacy Protection in AI-XR

In the literature, privacy-preserving techniques are broadly categorized into three classes: (i) cryptographic techniques, (ii) differential privacy, and (iii) federated and distributed ML, which are briefly discussed below.

5.1.1 Cryptographic Techniques. Cryptography refers to a practice of methodologies, aiming to construct and analyze communication protocols to ensure secure communication while achieving data integrity, authentication, non-repudiation, and data confidentiality. Generally, there are two common types of encryption methods: (i) symmetric encryption and (ii) asymmetric encryption method. The symmetric encryption method is a secret-key algorithm, in which the sender and receiver must share the same key to perform encryption and decryption of the data. Whereas, asymmetric encryption method (also known as public-key cryptography) uses two keys, i.e., public and private key, associated with an entity that requires to authenticate its identity electronically or encrypt data. The public key of each entity is published, whereas, the corresponding private key is always kept secret to perform encryption or decryption of data. In literature, **Ron Rivest, Adi Shamir, and Leonard Adleman (RSA)** [117], **Data Encryption Standard (DES)** [133], **Advanced Encryption Standard (AES)** [106], and **Secure Hash Algorithm (SHA)** [24] are a few commonly used algorithms used for data encryption. Different cryptographic techniques can be employed to convert readable information to an encrypted state, which can be later used at the receiver end after performing decryption. Below, we discuss some of the most commonly used encryption methods that can be used for the development of privacy-aware AI models.

Homomorphic Encryption. **Homomorphic encryption (HE)** is a computational approach that performs encryption while allowing computational tasks to be executed over encrypted data at the same time to ensure the privacy of the data. HE is defined as a public key cryptographic technique in which a pair of public and private keys is created to perform encryption and decryption operations on the data. The public key is used to encrypt the data before sharing it with the third party for further computational tasks including training and/or inference. Due to the homomorphic characteristics of this approach, the results can be decoded using the private key to visualize the results without showing them to third-party servers or unauthenticated users. In the ML literature, HE has been used for protecting the privacy of the users' data for different applications such as genome imputation [123], misinformation detection in text messages [6], and so on. Specifically, the AI models are trained and inferred using encrypted training and testing data, thus preserving the privacy of the sensitive data.

Secure Multi-party Computation. Secure multi-party computation (also known as secure computation) is a type of cryptographic technique that is focused on the development of collaborative methods to perform joint computation and calculate a function over joint inputs while possessing those inputs in an isolated fashion. Contrary to the traditional cryptographic methods, where cryptography ensures confidentiality and integrity of communication or storage, while the adversary is outside the system of users, this approach protects the users' privacy from each other while performing ML-based tasks, including training and inference activities.

Garbled Circuits. The idea of garbled circuits was first proposed by Yao in 1986 to perform two-party computation [169]. Garbled circuits can be used in a scenario where multiple parties are interested in performing some computation without sharing their data. Let us assume two parties (e.g., Alice and Bob, for the sake of simplicity) want to perform some computation using garbled circuits. Alice will send his input and function in the form of a garbled circuit, and Bob will utilize his garbled input with the garbled circuit to get the result of the required function, once he obtains

it from Alice in an oblivious fashion. In Reference [21], garbled circuits along with HE have been used to develop privacy-aware ML models, where the authors trained three classification models, namely, decision tree, Naïve Bayes, and hyperplane decision classifier, using the encrypted data.

Secret Sharing. In secret sharing, multiple parties collaborate in the computation by sharing their secrets among them while holding a "share" of the individual secrets. The secret can only be reconstructed by combining all the individual shares kept by participating parties, otherwise, it will be useless. In the literature, the secret-sharing technique has been successfully used for training AI models in a privacy-preserving way. For instance, Bonawitz et al. [20] used the secret-sharing technique to train an ML model by aggregating model updates from multiple parties in a privacy-aware way. In a similar study [67], authors used this technique for the development of a privacy-aware ML-based emotion recognition system leveraging client-server architecture. In their proposed framework, the secret-sharing technique was used for the communication of audio-visual data from the client side to the server, where an ensemble model based on a sparse autoencoder and a CNN model was used for the feature extraction from the collected data. The SVM classifier was then trained using the extracted features for the emotion recognition task. A secret-sharing-based parallelized variant of **principal component analysis (PCA)** for preserving data privacy is presented in Reference [19].

Secure Processors. Secure processors were pioneered by rogue software to protect sensitive code from being accessed by malicious actors at higher privilege levels. Secure processors are being used in different processors now to perform privacy-preserving operations, e.g., the Intel SGX processor. In Reference [101], SGX processors were used to develop a data-oblivious system for different ML techniques that include SVM, decision tree, matrix factorization, and k-mean clustering. The primary goal was to facilitate collaboration between multiple data proprietors performing the ML task on an SGX-empowered data center.

5.1.2   Differential Privacy. The idea of differential privacy is based on introducing noise in the data to protect sensitive information while ensuring the usefulness of the data [43]. Differential privacy is defined in terms of the task-specific concept of neighbor datasets, and it provides strong guarantees in ensuring the privacy of the data during algorithmic analysis [1]. Numerous differential privacy-based methods have been presented in the literature, such as **differentially private stochastic gradient descent (DP-SGD)** [42], **private-aggregation of teacher ensembles (PATE)** [174], exponential noise-based differential privacy-preserving methods to ensure privacy on large-scale data. These methods demonstrated better applicability in ML-based applications in various domains, including intelligent transportation services, smart/virtual personal assistants, and smart healthcare services.

5.1.3   Federated Learning. **Federated learning (FL)** refers to a distributed-ML paradigm that is capable of learning global ML models without directly accessing and/or exchanging data from edge devices. Intuitively, basic FL-based methods consist of a collaborative learning framework where each participant such as an edge device, network node, and local server can independently train a model using its local data. These edge devices then share their model parameters with a server, which then performs aggregation of the parameters after receiving parameter updates from each edge device. Finally, the server updates the parameters of the global model and shares the updated parameters with all participants. The iterative process of FL is continuous until the desired criteria as been fulfilled, e.g., validation accuracy/loss or the maximum number of communication rounds. In this way, a global model is trained without requiring the actual data from the FL participants. Subsequently, this sharing mechanism allows ML-based systems to learn from large-scale diverse data and develop a global model. Such methods can demonstrate better applicability in

terms of dealing with sensitive data in various human-centered applications such as AI-XR metaverse applications. Despite the success of FL in training an ML model with reliable performance while maintaining the privacy of the actual data, different attacks can be realized on the model being trained using the FL paradigm, e.g., backdoor attacks [160], label flipping attacks [110], free-riding attacks [89], poisoning attacks [5], and so on. Also, it has been demonstrated that sensitive information can be extracted from the shared parameters in FL settings [18].

5.2 Solutions to Combat Adversarial ML Attacks in AI-XR

In the literature, adversarially robust ML models have been mainly categorized into three categories [111]: (1) Data Modification; (2) Model Modification; and (3) Using Auxiliary Model. Moreover, a few methods leverage a hybrid approach in which multiple defensive techniques are used to develop adversarially robust ML models. Below, we discuss the most prominent methods in each category, and we refer readers interested in more details about these methods to comprehensive surveys that are focused on adversarial ML [4, 112, 150, 171].

5.2.1 Data Modification. Data modification methods modify the input data during the training or inference phase to mitigate the effects of adversarial noise. A few such famous methods are briefly described below.

— *Adversarial Re-training:* This method was proposed by Goodfellow et al. [54], and it is considered to be a basic method for mitigating the effect of adversarial perturbation in the trained model. In this method, adversarial examples are augmented in the training data, which is then used to (re)-train the model. This method has been extensively used in the literature, however, a few research studies demonstrated that the models trained using this method are not robust against multiple attacks [139].

— *Feature Squeezing:* Xu et al. [164] presented a feature-squeezing-based approach that aims to squeeze feature space of input that may be exploited in response to an adversary. In this regard, the heterogeneous feature vectors have been collectively joined into a single space to reduce available feature space. Although, the proposed defense method achieved significant performance against small perturbations. However, it was found less effective against iterative adversarial attacks [62].

— *Input Reconstruction:* Input reconstruction-based defense methods have been proposed to mitigate the effect of adversarial attacks. These methods transform adversarial examples into legitimate samples by cleaning adversarial noise using an appropriate technique, e.g., using an autoencoder [56].

5.2.2 Model Modification. Model modification methods aim at modifying the parameters of trained ML models to defuse the effect of adversarial attacks. Commonly used model modification methods are described below.

— *Gradient Regularization:* This method allows complex neural networks to bring a partial surge in training computational complexity to improve the performance of the network regardless of any prior knowledge about adversarial attacks. This idea was coined by Ross et al. [119] to improve the performance of CNN models on classification tasks. Though the proposed method achieved significant improvement in CNNs' robustness, it also increases the computational cost of models, which prejudices the performance in real-world ML-based applications.

— *Defensive Distillation:* Distillation in a neural network was initially conceptualized by Hinton et al. [65] to establish knowledge sharing from a larger network to a smaller one. Later, Papernot et al. [105] extended this notion by developing a distillation-based defense

mechanism against adversarial attacks, which is known as defensive distillation. In this method, the larger model is trained over hard labels to maximize accuracy while predicting the output probabilities of the baseline smaller model. This method is successful in mitigating the effect of small adversarial perturbation, and it fails in the presence of strong adversarial perturbations, e.g., adversarial examples generated using C&W attack [28].

— *Network Verification:* In this method, certain properties of the ML/DL model are verified, e.g., validating the output of models, produced in response to the corresponding input samples. Katz et al. [73] presented ReLU and **satisfiability modulo theory (SMT)**-based network verification method to make complex neural networks robust against adversarial examples. In a similar study, authors have proposed a scalable quantitative verification framework for DNNs to prove formal probabilistic property against adversarial attacks [14].

5.2.3 Using Auxiliary Model. Methods aiming to robustify ML models in this category use an additional model either to detect adversarial examples or to clean adversarial perturbations. A few methods are described below.

— *Adversarial Detection:* In such methods, a detector model is used to differentiate between normal and adversarial inputs, e.g., a binary classifier [90].
— *Ensembling Defenses:* In this defense strategy, an ensemble of different defensive techniques is created to withstand different adversarial attacks. PixelDefend is the most famous ensemble defense method, which consists of two defense approaches, i.e., input reconstruction and adversarial detection [132].
— *Using Generative Modeling:* These types of methods leverage different ML/DL-based generative models for cleaning adversarial noise in adversarial examples to project them back to the same data manifold.

5.3 Solutions for AI-XR Transparency and Trust Challenges

The true potential of AI-based applications in AI-XR metaverse applications can only be realized when they are developed using fine-grained personal data for making personalized recommendations and predictions, which is only possible when users fully trust the underlying system. Therefore, addressing the challenges related to the trustworthiness aspects of AI-XR metaverse applications is very important. From an AI perspective, trustworthiness itself requires predictability, interpretability, explainability, safety, and robustness. Below, we discuss different methods that can be used to accomplish trustworthiness in AI applications.

5.3.1 Explainable and Interpretable AI. An AI model is referred to as explainable if it can explain the ability of parameters to justify the results. Explainability aims at making an AI model understandable to humans. In recent years, substantial research efforts have been conducted to enhance explainability, trustworthiness, and interpretability in AI models. **Fairness, Accountability, and Transparency in Machine Learning (FAT-ML)** [47] and **Defense Advanced Research Projects Agency (DARPA)**, explainable AI program [59] are the two famous research groups working in this context. The literature argues that explainable models can be the first step toward converting black-box AI models into white-box models [2]. We refer interested readers to a comprehensive survey focused on explaining black-box models for a more detailed discussion of related terminologies and for a detailed classification of various explainable AI methods for a given type of black box system [58]. On the contrary, in Reference [120], a counter argument is provided where the author emphasized the need to develop inherently interpretable models instead of developing methods to explain black-box models (involving critical decisions).

Interpretable models refer to the models that explain themselves. In simple words, an AI model is said to be interpretable if its decision against some input is logically understandable such as which factors influenced the AI model to reach that decision. In the literature, various methods have been presented to leverage interpretability in ML models. These methods ensure that the predictions of interpretable models are unbiased, which ultimately makes it easier to trust these systems in human society. It is worth noting that the terms interpretable and explainable are interchangeably used in the literature, however, they are different in terms of domain-specific definitions; moreover, there is no exact definition of these terms [114]. In summary, explainable AI is more focused on explaining and communicating AI's output to humans, while interpretable AI is concerned with the understanding of the internal processing of AI models. A detailed taxonomy of different explainable and interpretable AI methods can be found in References [2, 114].

5.3.2 Trustworthy AI. The relevant literature emphasizes two famous sets of principles that can be used to attain trustworthy AI. One of them is developed by European Commission's AI **High-Level Expert Group (HLEG)** [66] and the other one is defined by **Organisation for Economic Co-operation and Development (OECD)** [170]. The following are the seven essential principles outlined in OECD: (1) Human agency and oversight; (2) Technical robustness and safety; (3) Privacy and data governance; (4) Transparency; (5) Diversity, non-discrimination, and fairness; (6) Environmental and societal well-being; and (7) Accountability.

Similarly, the following principles are outlined in HLEG: (1) Inclusive growth, sustainable development, and well-being; (2) Human-centered values and fairness; (3) Transparency and explainability; (4) Robustness, security, and safety; and (5) Accountability. One of the key noticeable insights from the above two principles set is that they mainly emphasized the explainability, security, fairness, safety, and robustness aspects of AI. Therefore, these are the essential requirements that need to be fulfilled to develop trustworthy AI-based applications. In addition, we can see that these principles are essentially human-centric and respect ethical norms. As potential AI-XR metaverse applications will be more human-focused, therefore, the above-mentioned principles can be leveraged to develop trustworthy AI-based applications for the metaverse.

5.4 Solutions for Ethical Challenges in AI-XR

5.4.1 Human in the Loop. The metaverse's inherent complexity raises different security issues. For instance, it can be envisioned that the metaverse administrators will have to push automation, that is, to handle more tasks with algorithms, rather than with human operators, due to the requirement of managing a large number of users, applications, and services. The generated data will be much larger than those managed by the current Web platforms. Delegating tasks to algorithms, especially those implemented with state-of-the-art AI approaches, is necessary to meet high-level efficiency and scalability. However, in the current social media and the Internet, we have even started to realize the implications of using algorithms for managing societally relevant tasks. Despite the significant performance, these algorithms suffer from various issues. Some authors writing on the governance of metaverse have proposed the use of a modular approach for the development of metaverse applications, as it allows adapting regulations to specific scenarios and then controlling the system accordingly [49].

5.4.2 Ethical and Responsible AI. To ensure socially desirable AI decisions, novel ways are required to be figured out to simultaneously minimize potential harms associated with the use of AI and its potential benefits. In this regard, the importance of taking an ethics-first approach towards the development of AI-based technologies becomes more plausible [52]. However, there are many challenges associated with the development of ethical AI pipelines due to distinct social norms and demographics of the human population, i.e., one ethical solution may be beneficial for a group

of people but it is highly possible that it will not be suitable for another group at the same time. Therefore, customized solutions are required to address such issues that can consider the social norms of target users while making AI-based decisions. In this regard, different ethical guidelines can be leveraged that can be potentially used for the development of pro-social AI solutions. The literature shows a groundswell of interest in ensuring ethical and responsible AI [107].

5.5 Situational Awareness

Situational awareness can be defined as the capacity to understand information perceived from the surrounding environment. The literature argues that situational awareness is a crucial and effective tool for monitoring the security of complex systems like metaverse [156]. Situational awareness can be used at the local and global levels for threat monitoring in a single metaverse or across multiple metaverses, respectively. The feasibility and potential of this tool have been extensively studied in the literature focused on XR and VR technology. For instance, Woodward et al. [156] performed a literature review that focused on the design of information presentation in AR headsets to enhance users' situational awareness. In Reference [72], authors performed immersive and realistic simulations to evaluate the effectiveness of audio-visual warning systems in increasing users' situational awareness in accident situations using VR and demonstrated that VR can assist drivers to remain alert.

5.6 Human-centric Approach for AI-XR Development

Metaverse is essentially a human-centric application [63]. To realize the real social impact of different AI-XR metaverse applications, they should be analyzed and developed using human-centric design thinking. Metaverse service providers and developers must pay attention to key stakeholders (i.e., humans) by prioritizing and considering their social norms, i.e., dignity, justice, and rights, and supporting goals including creativity, self-efficacy, social connections, and responsibility. The aforementioned characteristics can be inherited in AI-XR metaverse applications by following three key concepts proposed in Reference [131]. The first one is *Human-centric framework*—which guides developers and researchers to ensure human-centric thinking about high-level two-dimensional control. Second, *Design metaphors*—which point out how two key goals of AI and social norms are both valuable. However, the stakeholders such as developers, researchers, policymakers, and business leaders must combine them both in developing metaverse applications to provide ultimate benefits to the users. Third, *Governance Structures*—which ensure the bridge between the above-mentioned ethical principles and the practical measures needed to achieve the desired goals including reliable metaverse application development while ensuring cultural safety to increase privacy and trustworthiness of the users.

6 OPEN RESEARCH ISSUES

In this section, we highlight various open research issues that are particularly associated with the use of ML/DL models in different AI-XR metaverse applications.

6.1 Developing Generalizable Adversarial Defense Methods

Over the past few years, substantial research attention has been devoted to adversarial ML. However, the literature highlights that the attention devoted to developing adversarially robust ML/DL models is significantly less as compared to developing novel attack methodologies [109]. In the literature, different defensive techniques have been proposed to withstand adversarial ML attacks; however, each method only works in a specific setting and fails to withstand unseen and powerful attacks (consequently, fails to generalize across a wider class of attacks). However, the literature focused on adversarial ML shows that the diversity and severity of these attacks are increasing

with each passing day. Therefore, the development of hybrid and universal defensive techniques is the need of the hour. In addition, it is required that defense techniques should be developed while considering evolvable and adaptable adversaries (who can adapt their capabilities to break defense strategy). The threat of adversarial ML can be a major hurdle in the development of secure, safe, robust, and trustworthy AI-XR metaverse applications, and if it remains unaddressed, can cause unintended severe consequences to users and society. It is highly recommended to consider these aspects while developing ML/DL-empowered human-centric applications like the AI-XR metaverse. Moreover, the worst-case robustness test can be performed from an adversarial lens considering different attack surfaces in individual AI-XR metaverse application architecture.

6.2 Investigating Robustness of Privacy-preserving Methods

looseness-1 As discussed above, AI-XR metaverse applications will collect fine-grained data that may include personal attributes to provide personalized services (empowered by ML/DL models). The models trained with such data can be inferred to reconstruct privacy-related information that can be exploited to get intended outcomes and incentives. Although different privacy-preserving ML techniques proposed in the literature have been shown to be quite successful in preserving data privacy, the literature demonstrates that meaningful information can still be inferred even in the presence of an appropriate privacy-preserving method. For example, it has been demonstrated that homomorphic encryption (one of the widely used encryption techniques) is vulnerable to model extraction attacks [116]. Similarly, Boenisch et al. [18] showed that sensitive information can be reconstructed from the shared parameters in FL. This suggests that the investigation of vulnerabilities and limitations of existing privacy-preserving methods can be a good step toward developing robust privacy-preserving methods. Ideally, it is required that the ML/DL models should be developed in such a way that they are by design privacy-aware, i.e., they should not be able to learn any privacy-related features from the data that could be compromised upon model inferences.

6.3 Developing Generalizable Explainable and Interpretable Techniques

Another major limitation of DL models hindering their trustworthy applications in critical applications like the AI-XR metaverse is the lack of explainability and interpretability. This can also be exploited by adversarial agents to craft adversarial perturbations to realize attacks on different AI-XR metaverse applications. Although significant research interest has been devoted to the development of novel techniques to explain and interpret DL models, the literature shows that their application is limited to a certain data type or application [114]. While the AI-XR metaverse applications will have a complex architecture that will simultaneously use multi-modal data for making different intelligent decisions, existing explainability and interpretability techniques cannot be directly used for explaining and interpreting ML/DL-driven decisions. More work is required to create methods that can be generalized across different data types, models, and applications.

6.4 Developing Ethical Data Analysis Pipelines

Current ML/DL models are not capable of considering different ethical norms that are necessary for human-centric applications like AI-XR metaverse applications. On the contrary, these considerations are yet very important to maximize potential benefits and minimize associated harms to ensure safe, robust, and fair data analysis. In AI-XR metaverse applications, different ML/DL models will be trained using massively large data collected by humans and their interactions with the real and virtual universe. While there is no guarantee that the AI decision will be ethically committed, because the data used for model training might contain data bias that will eventually

result in biased decisions. Moreover, the outcomes of ML/DL models will be just a reflection of human behavior (including moral failures even if they are not intentionally committed). Therefore, to increase the trust of different stakeholders involved in AI-XR metaverse applications (particularly, end-users) and to provide them with a sense of safety, fairness, and accountability, it is highly desirable to develop novel techniques to ensure fair and ethical data analysis empowered by various ML/DL techniques.

6.5 Pushing AI on Edge: Embedded ML

One of the feasible approaches to preserve the privacy of AI-XR metaverse users will be to deploy ML/DL models on their smart devices, e.g., smartphones, AR/VR gadgets, tablets, and so on. By doing so, models can be developed and inferred on their devices without requiring to transmit data to a central cloud. We envision that various AI-XR metaverse applications will potentially adopt embedded ML or edge-enabled ML due to the proliferation of different enabling gadgets and smart devices. However, numerous challenges related to underlying hardware computing capabilities will arise when sufficiently large ML/DL models will be deployed on resource-constrained devices. Also, the literature argues that the research on enabling edge AI is at its early stages of development [108]. Therefore, it is worth investigating the feasibility and potential of deploying M/DL models on embedded devices to ultimately develop secure, private, and robust systems to provide personalized services in AI-XR metaverse applications. We refer interested readers to a recent survey on analyzing the notion of edge-enabled metaverse applications for a more comprehensive discussion on the topic and various challenges [163].

6.6 AI-XR Metaverse-specific Security Solutions

The future AI-XR metaverse will have a complex structure and will be a combination of various enabling (complex) technologies (that possess their associated challenges related to privacy and security, e.g., adversarial ML). Moreover, the massive connectivity of numerous entities (users, services providers, organizations, etc.) along with the decentralization will even worsen the enormity of security and privacy in AI-XR metaverse applications. Individual vulnerabilities associated with each technology can be exploited to realize a more powerful attack to halt or get control of some services or the entire metaverse. If such vulnerabilities are left unaddressed, then they will eventually lead to novel challenges, thus making it challenging to ensure the secure, safe, and robust operation of metaverse services. Therefore, it is very crucial to understand such challenges and develop customized defense solutions to protect AI-XR metaverse applications and services in general.

7 CONCLUSIONS

In this article, we have analyzed various security, privacy, and trustworthiness challenges associated with the use of different **machine learning (ML)** and **deep learning (DL)** techniques in **artificial intelligence and extended reality (AI-XR)** metaverse applications. Specifically, considering the layered architecture of the metaverse, we developed a pipeline and highlighted different potential ML/DL use cases along with identified various vulnerabilities associated with their application. Furthermore, we provided a comprehensive overview of these challenges and discussed potential solutions that could be used to overcome such issues. To accentuate the implications of adversarial threats, we designed a customized case study (considering a prospective AI-XR metaverse application) and analyzed its security and privacy aspects. Finally, we discussed various open research issues that require further investigation. We envision that our work on this crucial topic will provide a one-stop solution to interested researchers who aim to develop secure, robust, and trustworthy AI-XR applications.

REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. ACM, 308–318.

[2] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

[3] James Aiken and Sandra Scott-Hayward. 2019. Investigating adversarial attacks against network intrusion detection systems in SDNs. In *Proceedings of the IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN'19)*. IEEE, 1–7.

[4] Naveed Akhtar, Ajmal Mian, Navid Kardan, and Mubarak Shah. 2021. Advances in adversarial attacks and defenses in computer vision: A survey. *IEEE Access* 9 (2021), 155161–155196.

[5] Asad Ali, Inaam Ilahi, Adnan Qayyum, Ihab Mohammed, Ala Al-Fuqaha, and Junaid Qadir. 2021. Incentive-driven federated learning and associated security challenges: A systematic review. *TechRxiv* (2021).

[6] Hassan Ali, Rana Tallal Javed, Adnan Qayyum, Amer AlGhadhban, Meshari Alazmi, Ahmad Alzamil, Khaled Al-utaibi, and Junaid Qadir. 2022. SPAM-DaS: Secure and privacy-aware misinformation detection as a service. *TechRxiv* (2022).

[7] Hassan Ali, Faiq Khalid, Hammad Ali Tariq, Muhammad Abdullah Hanif, Rehan Ahmed, and Semeen Rehman. 2019. SSCNets: Robustifying DNNs using secure selective convolutional filters. *IEEE Des. Test* 37, 2 (2019), 58–65.

[8] Hassan Ali, Muhammad Suleman Khan, Ala Al-Fuqaha, and Junaid Qadir. 2022. Tamp-X: Attacking explainable natural language classifiers through tampered activations. *Comput. Secur.* (2022), 102791. DOI : https://doi.org/10.1016/j.cose.2022.102791

[9] Hassan Ali, Muhammad Suleman Khan, Amer AlGhadhban, Meshari Alazmi, Ahmad Alzamil, Khaled Al-Utaibi, and Junaid Qadir. 2021. All your fake detector are belong to us: Evaluating adversarial robustness of fake-news detectors under black-box settings. *IEEE Access* 9 (2021), 81678–81692.

[10] H. Ali, M. S. Khan, A. AlGhadhban, M. Alazmi, A. Alzamil, K. Al-utaibi, and J. Qadir. 2023. Condetect: Detecting adversarially perturbed natural language inputs to deep classifiers through holistic analysis. *Computers & Security* 132 (2023), 103367.

[11] Hassan Ali, Surya Nepal, Salil S. Kanhere, and Sanjay Jha. 2020. HaS-Nets: A heal and select mechanism to defend DNNs against backdoor attacks for data collection scenarios. *arXiv preprint arXiv:2012.07474* (2020).

[12] Madyan Alsenwi, Nguyen H. Tran, Mehdi Bennis, Shashi Raj Pandey, Anupam Kumar Bairagi, and Choong Seon Hong. 2021. Intelligent resource slicing for eMBB and URLLC coexistence in 5G and beyond: A deep reinforcement learning based approach. *IEEE Trans. Wirel. Commun.* 20, 7 (2021), 4585–4600.

[13] Anurag Arnab, Ondrej Miksik, and Philip H. S. Torr. 2018. On the robustness of semantic segmentation models to adversarial attacks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 888–897.

[14] Teodora Baluta, Zheng Leong Chua, Kuldeep S. Meel, and Prateek Saxena. 2021. Scalable quantitative verification for deep neural networks. In *Proceedings of the IEEE/ACM 43rd International Conference on Software Engineering (ICSE'21)*. 312–323.

[15] Elisa Bertino and Nayeem Islam. 2017. Botnets and internet of things security. *Computer* 50, 2 (2017), 76–79.

[16] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. 2013. Evasion attacks against machine learning at test time. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 387–402.

[17] Battista Biggio, B. Nelson, and P. Laskov. 2012. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Conference on Machine Learning*. ArXiv e-prints, 1807–1814.

[18] Franziska Boenisch, Adam Dziedzic, Roei Schuster, Ali Shahin Shamsabadi, Ilia Shumailov, and Nicolas Papernot. 2021. When the curious abandon honesty: Federated learning is not private. *arXiv preprint arXiv:2112.02918* (2021).

[19] Dan Bogdanov, Liina Kamm, Sven Laur, and Ville Sokk. 2018. Implementation and evaluation of an algorithm for cryptographically private principal component analysis on genomic data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 15, 5 (2018), 1427–1432.

[20] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. 2017. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. ACM, 1175–1191.

[21] Raphael Bost, Raluca Ada Popa, Stephen Tu, and Shafi Goldwasser. 2015. Machine learning classification over encrypted data. In *Proceedings of the Network and Distributed System Security Symposium (NDSS'15)*.

[22] Nicholas Boucher, Ilia Shumailov, Ross Anderson, and Nicolas Papernot. 2022. Bad characters: Imperceptible NLP attacks. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'22)*. IEEE, 1987–2004.

[23] Lauren Buck and Rachel McDonnell. 2022. Security and privacy in the metaverse: The threat of the digital human. In *Proceedings of the 1st Workshop on Novel Challenges of Safety, Security and Privacy in Extended Reality*.

[24] James H. Burrows. 1995. *Secure Hash Standard.* Technical Report. Department of Commerce, Washington, DC.

[25] Muhammad Atif Butt, Asad Masood Khattak, Sarmad Shafique, Bashir Hayat, Saima Abid, Ki-Il Kim, Muhammad Waqas Ayub, Ahthasham Sajid, and Awais Adnan. 2021. Convolutional neural network based vehicle classification in adverse illuminous conditions for intelligent transportation systems. *Complexity* 2021 (2021), Article ID: 6644861.

[26] Muhammad Atif Butt and Faisal Riaz. 2022. CARL-D: A vision benchmark suite and large scale dataset for vehicle detection and scene segmentation. *Sig. Process.: Image Commun.* 104 (2022), 116667.

[27] Alberto Cannavo and Fabrizio Lamberti. 2020. How blockchain, virtual reality, and augmented reality are converging, and why. *IEEE Consum. Electron. Mag.* 10, 5 (2020), 6–13.

[28] Nicholas Carlini and David Wagner. 2017. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security.* 3–14.

[29] Peter Casey, Ibrahim Baggili, and Ananya Yarramreddy. 2019. Immersive virtual reality attacks and the human joystick. *IEEE Trans. Depend. Sec. Comput.* 18, 2 (2019), 550–562.

[30] Efstathios Chatzikyriakidis, Christos Papaioannidis, and Ioannis Pitas. 2019. Adversarial face de-identification. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'19).* IEEE, 684–688.

[31] Dawei Chen, Dan Wang, Yifei Zhu, and Zhu Han. 2021. Digital twin for federated analytics using a Bayesian approach. *IEEE Internet Things J.* 8, 22 (2021), 16301–16312.

[32] Mingzhe Chen, Ursula Challita, Walid Saad, Changchuan Yin, and Mérouane Debbah. 2019. Artificial neural networks-based machine learning for wireless networks: A tutorial. *IEEE Commun. Surv. Tutor.* 21, 4 (2019), 3039–3071.

[33] Xiaoyi Chen, Ahmed Salem, Michael Backes, Shiqing Ma, and Yang Zhang. 2021. BadNL: Backdoor attacks against NLP models. In *Proceedings of the ICML Workshop on Adversarial Machine Learning.*

[34] Yanjiao Chen, Xueluan Gong, Qian Wang, Xing Di, and Huayang Huang. 2020. Backdoor attacks and defenses for deep neural networks in outsourced cloud environments. *IEEE Netw.* 34, 5 (2020), 141–147.

[35] Ruizhi Cheng, Nan Wu, Songqing Chen, and Bo Han. 2022. Will metaverse be NextG internet? Vision, hype, and reality. *arXiv preprint arXiv:2201.12894* (2022).

[36] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 3213–3223.

[37] Ali Dabouei, Sobhan Soleymani, Jeremy Dawson, and Nasser Nasrabadi. 2019. Fast geometrically-perturbed adversarial faces. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV'19).* IEEE, 1979–1988.

[38] Jaybie A. De Guzman, Kanchana Thilakarathna, and Aruna Seneviratne. 2019. Security and privacy approaches in mixed reality: A literature survey. *ACM Comput. Surv.* 52, 6 (2019), 1–37.

[39] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, 248–255.

[40] Roberto Di Pietro and Stefano Cresci. 2021. Metaverse: Security and privacy issues. In *Proceedings of the 3rd IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA'21).* IEEE, 281–288.

[41] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. 2019. Efficient decision-based black-box adversarial attacks on face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 7714–7722.

[42] Jian Du, Song Li, Moran Feng, and Siheng Chen. 2021. Dynamic differential-privacy preserving SGD. *arXiv:2111.00173* (2021).

[43] Cynthia Dwork. 2011. Differential privacy. *Encyc. Cryptog. Secur.* (2011), 338–340.

[44] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2017. HotFlip: White-box adversarial examples for text classification. *arXiv preprint arXiv:1712.06751* (2017).

[45] Ben Falchuk, Shoshana Loeb, and Ralph Neff. 2018. The social metaverse: Battle for privacy. *IEEE Technol. Soc. Mag.* 37, 2 (2018), 52–61.

[46] Saeed Banaeian Far and Azadeh Imani Rad. 2022. Applying digital twins in metaverse: User interface, security and privacy challenges. *J. Metaverse* 2, 1 (2022), 8–16.

[47] M. L. FAT. 2018. Fairness, accountability, and transparency in machine learning. Retrieved August 22, 2023 from https://www.fatml.org/

[48] Yu Feng, Benteng Ma, Jing Zhang, Shanshan Zhao, Yong Xia, and Dacheng Tao. 2022. FIBA: Frequency-injection based backdoor attack in medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 20876–20885.

[49] Carlos Bermejo Fernandez and Pan Hui. 2022. Life, the metaverse and everything: An overview of privacy, ethics, and governance in metaverse. *arXiv preprint arXiv:2204.01480* (2022).

[50] Volker Fischer, Mummadi Chaithanya Kumar, Jan Hendrik Metzen, and Thomas Brox. 2017. Adversarial examples for semantic image segmentation. *arXiv preprint arXiv:1703.01101* (2017).

[51] Martin Flintham, Christian Karner, Khaled Bachour, Helen Creswick, Neha Gupta, and Stuart Moran. 2018. Falling for fake news: Investigating the consumption of news via social media. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, 376.

[52] Luciano Floridi. 2021. *Ethics, Governance, and Policies in Artificial Intelligence*. Springer.

[53] Siddhant Garg and Goutham Ramakrishnan. 2020. Bae: BERT-based adversarial examples for text classification. *arXiv preprint arXiv:2004.01970* (2020).

[54] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[55] Gaurav Goswami, Nalini Ratha, Akshay Agarwal, Richa Singh, and Mayank Vatsa. 2018. Unravelling robustness of deep learning based face recognition against adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[56] Shixiang Gu and Luca Rigazio. 2014. Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068* (2014).

[57] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. 2019. BadNets: Evaluating backdooring attacks on deep neural networks. *IEEE Access* 7 (2019), 47230–47244.

[58] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Comput. Surv.* 51, 5 (2018), 1–42.

[59] David Gunning. 2017. Explainable artificial intelligence (XAI). Defense Advanced Research Projects Agency (DARPA), Retrieved August 22, 2023 from https://www.darpa.mil/program/explainable-artificialintelligence

[60] Osama Halabi, Shidin Balakrishnan, Sarada Prasad Dakua, Nassir Navab, and Mohammed Warfa. 2020. Virtual and augmented reality in surgery. In *The Disruptive Fourth Industrial Revolution*. Springer, 257–285. DOI : https://doi.org/10.1007/978-3-030-48230-5_11

[61] Abdullah Hamdi, Sara Rojas, Ali Thabet, and Bernard Ghanem. 2020. AdvPC: Transferable adversarial perturbations on 3D point clouds. In *Proceedings of the European Conference on Computer Vision*. Springer, 241–257.

[62] Warren He, James Wei, Xinyun Chen, Nicholas Carlini, and Dawn Song. 2017. Adversarial example defense: Ensembles of weak defenses are not strong. In *Proceedings of the 11th USENIX Workshop on Offensive Technologies (WOOT'17)*.

[63] Lynne Heller and Lizbeth Goodman. 2016. What do avatars want now? Posthuman embodiment and the technological sublime. In *Proceedings of the 22nd International Conference on Virtual System & Multimedia (VSMM'16)*. IEEE, 1–4.

[64] Jan Hendrik Metzen, Mummadi Chaithanya Kumar, Thomas Brox, and Volker Fischer. 2017. Universal adversarial perturbations against semantic image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17)*. IEEE, 2755–2764.

[65] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv:1503.02531* (2015).

[66] Andreas Holzinger, Matthias Dehmer, Frank Emmert-Streib, Rita Cucchiara, Isabelle Augenstein, Javier Del Ser, Wojciech Samek, Igor Jurisica, and Natalia Díaz-Rodríguez. 2022. Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. *Inf. Fusion* 79 (2022), 263–278.

[67] M. Shamim Hossain and Ghulam Muhammad. 2019. Emotion recognition using secure edge and cloud computing. *Inf. Sci.* 504 (2019), 589–601.

[68] Thien Huynh-The, Quoc-Viet Pham, Xuan-Qui Pham, Thanh Thi Nguyen, Zhu Han, and Dong-Seong Kim. 2022. Artificial intelligence for the metaverse: A survey. *arXiv preprint arXiv:2202.10336* (2022).

[69] Olakunle Ibitoye, Rana Abou-Khamis, Ashraf Matrawy, and M. Omair Shafiq. 2019. The threat of adversarial attacks on machine learning in network security—A survey. *arXiv preprint arXiv:1911.02621* (2019).

[70] Saquib Irtiza, Latifur Khan, and Kevin W. Hamlen. 2022. SentMod: Hidden backdoor attack on unstructured textual data. In *Proceedings of the IEEE 8th International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing,(HPSC) and IEEE International Conference on Intelligent Data and Security (IDS)*. IEEE, 224–231.

[71] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 8018–8025.

[72] Uijong Ju, Lewis L. Chuang, and Christian Wallraven. 2020. Acoustic cues increase situational awareness in accident situations: A VR car-driving study. *IEEE Trans. Intell. Transport. Syst.* (2020).

[73] Guy Katz, Clark Barrett, David L. Dill, Kyle Julian, and Mykel J. Kochenderfer. 2017. Reluplex: An efficient SMT solver for verifying deep neural networks. In *Proceedings of the International Conference on Computer Aided Verification.* Springer, 97–117.

[74] Manish Kesarwani, Bhaskar Mukhoty, Vijay Arya, and Sameep Mehta. 2018. Model extraction warning in MLaaS paradigm. In *Proceedings of the 34th Annual Computer Security Applications Conference.* ACM, 371–380.

[75] Faiq Khalid, Hassan Ali, Muhammad Abdullah Hanif, Semeen Rehman, Rehan Ahmed, and Muhammad Shafique. 2020. FaDec: A fast decision-based attack for adversarial machine learning. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'20).* IEEE, 1–8.

[76] Faiq Khalid, Hassan Ali, Hammad Tariq, Muhammad Abdullah Hanif, Semeen Rehman, Rehan Ahmed, and Muhammad Shafique. 2019. QuSecNets: Quantization-based defense mechanism for securing deep neural network against adversarial attacks. In *Proceedings of the IEEE 25th International Symposium on On-Line Testing and Robust System Design (IOLTS'19).* IEEE, 182–187.

[77] Pınar Kürtünlüoğlu, Beste Akdik, and Enis Karaarslan. 2022. Security of virtual reality authentication methods in metaverse: An overview. *arXiv preprint arXiv:2209.06447* (2022).

[78] Jaron Lanier. 2018. *Ten Arguments for Deleting Your Social Media Accounts Right Now.* Random House.

[79] Siddique Latif, Adnan Qayyum, Muhammad Usama, Junaid Qadir, Andrej Zwitter, and Muhammad Shahzad. 2019. Caveat emptor: The risks of using big data for human development. *IEEE Technol. Soc. Mag.* 38, 3 (2019), 82–90.

[80] John D. Lee and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. *Hum. Fact.* 46, 1 (2004), 50–80.

[81] Lik-Hang Lee, Tristan Braud, Pengyuan Zhou, Lin Wang, Dianlei Xu, Zijun Lin, Abhishek Kumar, Carlos Bermejo, and Pan Hui. 2021. All one needs to know about metaverse: A complete survey on technological singularity, virtual ecosystem, and research agenda. *arXiv preprint arXiv:2110.05352* (2021).

[82] Mark Lee and Zico Kolter. 2019. On physical adversarial patches for object detection. *arXiv preprint arXiv:1906.11897* (2019).

[83] Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2018. Textbugger: Generating adversarial text against real-world applications. *arXiv preprint arXiv:1812.05271* (2018).

[84] Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-attack: Adversarial attack against BERT using BERT. *arXiv preprint arXiv:2004.09984* (2020).

[85] Xinke Li, Zhirui Chen, Yue Zhao, Zekun Tong, Yabang Zhao, Andrew Lim, and Joey Tianyi Zhou. 2021. PointBA: Towards backdoor attacks in 3D point cloud. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'21).* 16492–16501.

[86] Yiming Li, Yanjie Li, Yalei Lv, Yong Jiang, and Shu-Tao Xia. 2021. Hidden backdoor attack against semantic segmentation models. *arXiv preprint arXiv:2103.04038* (2021).

[87] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing design practices for explainable AI user experiences. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.* 1–15.

[88] Q. Vera Liao and Kush R. Varshney. 2021. Human-centered explainable ai (XAI): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790* (2021).

[89] Jierui Lin, Min Du, and Jian Liu. 2019. Free-riders in federated learning: Attacks and defenses. *arXiv preprint arXiv:1911.12560* (2019).

[90] Jiajun Lu, Theerasit Issaranon, and David Forsyth. 2017. SafetyNet: Detecting and rejecting adversarial examples robustly. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17).* 446–454.

[91] Changqing Luo, Jinlong Ji, Qianlong Wang, Xuhui Chen, and Pan Li. 2018. Channel state information prediction for 5G wireless communications: A deep learning approach. *IEEE Trans. Netw. Sci. Eng.* 7, 1 (2018), 227–236.

[92] Bernard Marr. 2021. *Extended Reality in Practice.* Wiley.

[93] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).

[94] Enrico Meloni, Matteo Tiezzi, Luca Pasqualini, Marco Gori, and Stefano Melacci. 2021. Messing up 3D virtual environments: Transferable adversarial 3D objects. In *Proceedings of the 20th IEEE International Conference on Machine Learning and Applications.* IEEE, 1–8.

[95] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* 267 (2019), 1–38.

[96] Luis Muñoz-González, Battista Biggio, Ambra Demontis, Andrea Paudice, Vasin Wongrassamee, Emil C. Lupu, and Fabio Roli. 2017. Towards poisoning of deep learning algorithms with back-gradient optimization. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security.* 27–38.

[97] Cong T. Nguyen, Dinh Thai Hoang, Diep N. Nguyen, and Eryk Dutkiewicz. 2021. Metachain: A novel blockchain-based framework for metaverse applications. *arXiv preprint arXiv:2201.00759* (2021).

[98] Huansheng Ning, Hang Wang, Yujia Lin, Wenxi Wang, Sahraoui Dhelim, Fadi Farha, Jianguo Ding, and Mahmoud Daneshmand. 2021. A survey on metaverse: The state-of-the-art, technologies, applications, and challenges. *arXiv preprint arXiv:2111.09673* (2021).

[99] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, Ioannis Kompatsiaris, Katharina Kinder-Kurlanda, Claudia Wagner, Fariba Karimi, Miriam Fernandez, Harith Alani, Bettina Berendt, Tina Kruegel, Christian Heinze, Klaus Broelemann, Gjergji Kasneci, Thanassis Tiropanis, and Steffen Staab. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisc. Rev.: Data Min. Knowl. Discov.* 10, 3 (2020), e1356.

[100] Immersive and Addictive Technologies, UK House of Commons DCMS Committee, UK Parliament. 2019. Retrieved August 22, 2023 from https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1846/1846.pdf

[101] Olga Ohrimenko, Felix Schuster, Cédric Fournet, Aastha Mehta, Sebastian Nowozin, Kapil Vaswani, and Manuel Costa. 2016. Oblivious multi-party machine learning on trusted processors. In *Proceedings of the 25th USENIX Security Symposium (USENIX Security'16)*. 619–636.

[102] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Front. Big Data* 2 (2019), 13.

[103] Daniel W. Otter, Julian R. Medina, and Jugal K. Kalita. 2020. A survey of the usages of deep learning for natural language processing. *IEEE Trans. Neural Netw. Learn. Syst.* 32, 2 (2020), 604–624.

[104] Xudong Pan, Mi Zhang, Beina Sheng, Jiaming Zhu, and Min Yang. 2022. Hidden trigger backdoor attack on NLP models via linguistic style manipulation. In *Proceedings of the 31st USENIX Security Symposium (USENIX Security'22)*. 3611–3628.

[105] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. 2016. Distillation as a defense to adversarial perturbations against deep neural networks. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'16)*. IEEE, 582–597.

[106] NIST FIPS Pub. 2001. 197: Advanced encryption standard (AES). *Fed. Inf. Process. Stand. Pub.* 197, 441 (2001).

[107] Junaid Qadir, Mohammad Qamar Islam, and Ala Al-Fuqaha. 2022. Toward accountable human-centered AI: Rationale and promising directions. *J. Inf., Commun. Ethics Soc.* 20, 2 (2022).

[108] Adnan Qayyum, Kashif Ahmad, Muhammad Ahtazaz Ahsan, Ala Al-Fuqaha, and Junaid Qadir. 2022. Collaborative federated learning for healthcare: Multi-modal COVID-19 diagnosis at the edge. *IEEE Open J. Comput. Soc.* 3 (2022).

[109] Adnan Qayyum, Aneeqa Ijaz, Muhammad Usama, Waleed Iqbal, Junaid Qadir, Yehia Elkhatib, and Ala Al-Fuqaha. 2020. Securing machine learning in the cloud: A systematic review of cloud machine learning security. *Front. Big Data* 3 (2020), 587139.

[110] Adnan Qayyum, Muhammad Umar Janjua, and Junaid Qadir. 2022. Making federated learning robust to adversarial attacks by learning data and model association. *Comput. Secur.* 121 (2022), 102827.

[111] Adnan Qayyum, Junaid Qadir, Muhammad Bilal, and Ala Al-Fuqaha. 2020. Secure and robust machine learning for healthcare: A survey. *IEEE Rev. Biomed. Eng.* 14 (2020), 156–180.

[112] Adnan Qayyum, Muhammad Usama, Junaid Qadir, and Ala Al-Fuqaha. 2020. Securing connected & autonomous vehicles: Challenges posed by adversarial machine learning and the way forward. *IEEE Commun. Surv. Tutor.* 22, 2 (2020), 998–1026.

[113] Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. 2019. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *Proceedings of the International Conference on Machine Learning*. PMLR, 5231–5240.

[114] Khansa Rasheed, Adnan Qayyum, Mohammed Ghaly, Ala Al-Fuqaha, Adeel Razi, and Junaid Qadir. 2022. Explainable, trustworthy, and ethical machine learning for healthcare: A survey. *Comput. Biol. Med.* 149 (2022), 106043.

[115] Dirk Reiners, Mohammad Reza Davahli, Waldemar Karwowski, and Carolina Cruz-Neira. 2021. The combination of artificial intelligence and extended reality: A systematic review. *Front. Virt. Real.* 2 (2021), 114.

[116] Robert Nikolai Reith, Thomas Schneider, and Oleksandr Tkachenko. 2019. Efficiently stealing your machine learning models. In *Proceedings of the 18th ACM Workshop on Privacy in the Electronic Society*. 198–210.

[117] Ronald L. Rivest, Adi Shamir, and Leonard Adleman. 1978. A method for obtaining digital signatures and public-key cryptosystems. *Commun. ACM* 21, 2 (1978), 120–126.

[118] Louis B. Rosenberg. 2022. Regulation of the metaverse: A roadmap. In *Proceedings of the 6th International Conference on Virtual and Augmented Reality Simulations (ICVARS'22)*.

[119] Andrew Slavin Ross and Finale Doshi-Velez. 2018. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.

[120] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1, 5 (2019), 206–215.

[121] Meysam Sadeghi and Erik G. Larsson. 2018. Adversarial attacks on deep-learning based radio signal classification. *IEEE Wirel. Commun. Lett.* 8, 1 (2018), 213–216.

[122] Yalin E. Sagduyu, Yi Shi, and Tugba Erpek. 2019. IoT network security from the perspective of adversarial deep learning. In *Proceedings of the 16th Annual International Conference on Sensing, Communication, and Networking*. IEEE, 1–9.

[123] Esha Sarkar, Eduardo Chielle, Gamze Gürsoy, Oleg Mazonka, Mark Gerstein, and Michail Maniatakos. 2021. Fast and scalable private genotype imputation using machine learning and partially homomorphic encryption. *IEEE Access* 9 (2021), 93097–93110.

[124] Tegjyot Singh Sethi and Mehmed Kantardzic. 2018. Data driven exploratory attacks on black box classifiers in adversarial domains. *Neurocomputing* 289 (2018), 129–143.

[125] Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. 2018. Poison frogs! Targeted clean-label poisoning attacks on neural networks. *Adv. Neural Inf. Process. Syst.* 31 (2018).

[126] Kyarash Shahriari and Mana Shahriari. 2017. IEEE standard review—Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. In *Proceedings of the IEEE Canada International Humanitarian Technology Conference*. IEEE, 197–201.

[127] Jiacheng Shang, Si Chen, Jie Wu, and Shu Yin. 2020. ARSpy: Breaking location-based multi-player augmented reality application for user location tracking. *IEEE Trans. Mob. Comput.* 21, 2 (2020).

[128] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. 2016. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. 1528–1540.

[129] Changyang She, Rui Dong, Zhouyou Gu, Zhanwei Hou, Yonghui Li, Wibowo Hardjawana, Chenyang Yang, Lingyang Song, and Branka Vucetic. 2020. Deep learning for ultra-reliable and low-latency communications in 6G networks. *IEEE Netw.* 34, 5 (2020), 219–225.

[130] Meng Shen, Zelin Liao, Liehuang Zhu, Ke Xu, and Xiaojiang Du. 2019. VLA: A practical visible light-based attack on face recognition systems in physical world. *Proc. ACM Interact., Mob., Wear. Ubiq. Technol.* 3, 3 (2019), 1–19.

[131] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *Int. J. Hum.–comput. Interact.* 36, 6 (2020), 495–504.

[132] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. 2018. PixelDefend: Leveraging generative models to understand and defend against adversarial examples. In *Proceedings of the International Conference on Learning Representations (ICLR'18)*.

[133] Data Encryption Standard et al. 1999. Data encryption standard. *Fed. Inf. Process. Stand. Pub.* 112 (1999).

[134] Jacob Steinhardt, Pang Wei W. Koh, and Percy S. Liang. 2017. Certified defenses for data poisoning attacks. *Adv. Neural Inf. Process. Syst.* 30 (2017).

[135] Harini Suresh and John V. Guttag. 2019. A framework for understanding unintended consequences of machine learning. *arXiv* (2019).

[136] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[137] Sudeep Tanwar, Qasim Bhatia, Pruthvi Patel, Aparna Kumari, Pradeep Kumar Singh, and Wei-Chiang Hong. 2019. Machine learning adoption in blockchain-based smart applications: The challenges, and a way forward. *IEEE Access* 8 (2019), 474–488.

[138] Fei Tao, He Zhang, Ang Liu, and Andrew Y. C. Nee. 2018. Digital twin in industry: State-of-the-art. *IEEE Trans. Industr. Inform.* 15, 4 (2018), 2405–2415.

[139] Florian Tramer and Dan Boneh. 2019. Adversarial training and robustness for multiple perturbations. *Adv. Neural Inf. Process. Syst.* 32 (2019).

[140] Godwin Brown Tunze, Thien Huynh-The, Jae-Min Lee, and Dong-Seong Kim. 2020. Sparsely connected CNN for efficient automatic modulation recognition. *IEEE Trans. Vehic. Technol.* 69, 12 (2020), 15557–15568.

[141] Muhammad Usama, Muhammad Asim, Siddique Latif, Junaid Qadir, and Ala-Al-Fuqaha. 2019. Generative adversarial networks for launching and thwarting adversarial attacks on network intrusion detection systems. In *Proceedings of the 15th International Wireless Communications & Mobile Computing Conference (IWCMC'19)*. IEEE, 78–83.

[142] Muhammad Usama, Inaam Ilahi, Junaid Qadir, Rupendra Nath Mitra, and Mahesh K. Marina. 2021. Examining machine learning for 5G and beyond through an adversarial lens. *IEEE Internet Comput.* 25, 2 (2021), 26–34.

[143] Muhammad Usama, Junaid Qadir, and Ala Al-Fuqaha. 2018. Adversarial attacks on cognitive self-organizing networks: The challenge and the way forward. In *Proceedings of the IEEE 43rd Conference on Local Computer Networks Workshops (LCN Workshops'18)*. IEEE, 90–97.

[144] Muhammad Usama, Junaid Qadir, and Ala Al-Fuqaha. 2019. Black-box adversarial ML attack on modulation classification. *arXiv* (2019).

[145] Salah-ud-din Farooq, Muhammad Usama, Junaid Qadir, and Muhammad Ali Imran. 2019. Adversarial ML attack on self organizing cellular networks. In *Proceedings of the UK/China Emerging Technologies (UCET'19)*. IEEE, 1–5.

[146] Shannon Vallor. 2016. *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford University Press.

[147] Samaikya Valluripally, Aniket Gulhane, Reshmi Mitra, Khaza Anuarul Hoque, and Prasad Calyam. 2020. Attack trees for security and privacy in social virtual reality learning environments. In *Proceedings of the 17th Annual Consumer Communications & Networking Conference*. IEEE, 1–9.

[148] Xupeng Wang, Mumuxin Cai, Ferdous Sohel, Nan Sang, and Zhengwei Chang. 2021. Adversarial point cloud perturbations against 3D object detection in autonomous driving systems. *Neurocomputing* 466 (2021), 27–36.

[149] Yuntao Wang, Zhou Su, Jianbing Ni, Ning Zhang, and Xuemin Shen. 2021. Blockchain-empowered space-air-ground integrated networks: Opportunities, challenges, and solutions. *IEEE Commun. Surv. Tutor.* 24, 1 (2021), 160–209.

[150] Yuntao Wang, Zhou Su, Ning Zhang, Rui Xing, Dongxiao Liu, Tom H. Luan, and Xuemin Shen. 2022. A survey on metaverse: Fundamentals, security, and privacy. *IEEE Commun. Surv. Tutor.* 25, 1 (2022).

[151] Yajie Wang, Yu-an Tan, Wenjiao Zhang, Yuhang Zhao, and Xiaohui Kuang. 2020. An adversarial attack on DNN-based black-box object detectors. *J. Netw. Comput. Applic.* 161 (2020), 102634.

[152] Segev Wasserkrug, Avigdor Gal, and Opher Etzion. 2008. Inference of security hazards from event composition based on incomplete or uncertain information. *IEEE Trans. Knowl. Data Eng.* 20, 8 (2008), 1111–1114.

[153] Jianhao Wei, Junyi Li, Yaping Lin, and Jin Zhang. 2020. LDP-based social content protection for trending topic recommendation. *IEEE Internet Things J.* 8, 6 (2020), 4353–4372.

[154] Xingxing Wei, Siyuan Liang, Ning Chen, and Xiaochun Cao. 2018. Transferable adversarial attacks for image and video object detection. *arXiv preprint arXiv:1811.12641* (2018).

[155] Emily Wenger, Josephine Passananti, Arjun Nitin Bhagoji, Yuanshun Yao, Haitao Zheng, and Ben Y. Zhao. 2021. Backdoor attacks against deep learning systems in the physical world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6206–6215.

[156] Julia Woodward and Jaime Ruiz. 2022. Analytic review of using augmented reality for situational awareness. *IEEE Trans. Visualiz. Comput. Graph.* 29, 4 (2022).

[157] Tong Wu, Tianhao Wang, Vikash Sehwag, Saeed Mahloujifar, and Prateek Mittal. 2022. Just rotate it: Deploying backdoor attacks via rotation transformation. *arXiv preprint arXiv:2207.10825* (2022).

[158] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. 2021. TediGAN: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2256–2265.

[159] Chong Xiang, Charles R. Qi, and Bo Li. 2019. Generating 3D adversarial point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9136–9144.

[160] Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. 2019. DBA: Distributed backdoor attacks against federated learning. In *Proceedings of the International Conference on Learning Representations*.

[161] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. 2017. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17)*. 1369–1378.

[162] Yi Xie, Cong Shi, Zhuohang Li, Jian Liu, Yingying Chen, and Bo Yuan. 2020. Real-time, universal, and robust adversarial attacks against speaker recognition systems. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'20)*. IEEE, 1738–1742.

[163] Minrui Xu, Wei Chong Ng, Wei Yang Bryan Lim, Jiawen Kang, Zehui Xiong, Dusit Niyato, Qiang Yang, Xuemin Sherman Shen, and Chunyan Miao. 2022. A full dive into realizing the edge-enabled metaverse: Visions, enabling technologies, and challenges. *arXiv preprint arXiv:2203.05471* (2022).

[164] Weilin Xu, David Evans, and Yanjun Qi. 2017. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155* (2017).

[165] Weilin Xu, Yanjun Qi, and David Evans. 2016. Automatically evading classifiers. In *Proceedings of the Network and Distributed Systems Symposium*.

[166] Mingfu Xue, Can He, Jian Wang, and Weiqiang Liu. 2021. Backdoors hidden in facial features: A novel invisible backdoor attack against face recognition systems. *Peer-to-Peer Netw. Applic.* 14, 3 (2021), 1458–1474.

[167] Qinglin Yang, Yetong Zhao, Huawei Huang, Zehui Xiong, Jiawen Kang, and Zibin Zheng. 2022. Fusing blockchain and AI with metaverse: A survey. *IEEE Open J. Comput. Soc.* 3 (2022), 122–136.

[168] Ziqi Yang, Jiyi Zhang, Ee-Chien Chang, and Zhenkai Liang. 2019. Neural network inversion in adversarial setting via background knowledge alignment. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. 225–240.

[169] Andrew Chi-Chih Yao. 1986. How to generate and exchange secrets. In *Proceedings of the 27th Annual Symposium on Foundations of Computer Science (SFCS'86)*. IEEE, 162–167.

[170] Karen Yeung. 2020. Recommendation of the council on artificial intelligence (OECD). *Int. Legal Mater.* 59, 1 (2020), 27–34.

[171] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. 2019. Adversarial examples: Attacks and defenses for deep learning. *IEEE Trans. Neural Netw. Learn. Syst.* 30, 9 (2019).

[172] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Adv. Neural Inf. Process. Syst.* 32 (2019).

[173] Hantao Zhang, Wengang Zhou, and Houqiang Li. 2020. Contextual adversarial attacks for object detection. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'20)*. IEEE, 1–6.

[174] Qiuchen Zhang, Jing Ma, Jian Lou, Li Xiong, and Xiaoqian Jiang. 2020. Towards training robust private aggregation of teacher ensembles under noisy labels. In *Proceedings of the IEEE International Conference on Big Data (Big Data'20)*. IEEE, 1103–1110.

[175] Xinze Zhang, Junzhe Zhang, Zhenhua Chen, and Kun He. 2021. Crafting adversarial examples for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. 1967–1977.

[176] Ruoyu Zhao, Yushu Zhang, Youwen Zhu, Rushi Lan, and Zhongyun Hua. 2022. Metaverse: Security and privacy concerns. *arXiv preprint arXiv:2203.03854* (2022).

[177] Yaoyao Zhong and Weihong Deng. 2020. Towards transferable adversarial attack against deep face recognition. *IEEE Trans. Inf. Forens. Secur.* 16 (2020), 1452–1466.