Annals of Telecommunications
https://doi.org/10.1007/s12243-024-01031-7

Guarding 6G use cases: a deep dive into AI/ML threats in All-Senses meeting

Leyli Karaçay[1] · Zakaria Laaroussi[2] · Sonika ujjwal[2] · Elif Ustundag Soykan[3]

Received: 30 December 2023 / Accepted: 25 March 2024

Abstract
With the recent advances in 5G and 6G communications and the increasing need for immersive interactions due to pandemic, new use cases such as All-Senses meeting are emerging. To realize these use cases, numerous sensors, actuators, and virtual reality devices are used. Additionally, artificial intelligence (AI) and machine learning (ML) including generative AI can be used to analyze large amount of data generated by 6G networks and devices to enable new applications and services. While AI/ML technologies are evolving, they do not have the same level of security as well-known information technology components. So, AI/ML threats and their impacts can be overlooked. On the other hand, due to inherent characteristics of AI/ML components and design of AI/ML pipeline, AI/ML services can be a target for sophisticated attacks. In order to provide a holistic security view, the effect of AI/ML components should be investigated, threats should be identified, and countermeasures should be planned. Therefore, in this study, which is an extended version of our recent study (Karaçay et al. 2023), we shed the light on the use of AI/ML services including generative large language model scenarios in All-Senses meeting use case and their security aspects by carrying out a threat modeling using the STRIDE framework and attack tree methodology. Additionally, we point out some countermeasures for identified threats.

Keywords 6G · AI/ML · Threat analysis · Security · Holographic communication

1 Introduction

The COVID-19 pandemic has disrupted the way people used to work in office setups and introduced a new way of working from home. The same change happened in the online learning space as well. This change has been made possible by a number of video conference applications like Zoom, Google Meet, and Microsoft Teams which worked as enablers for holding different types of meetings, such as educational or work meetings, in an efficient and scalable manner. Now, imagine being able to touch and feel your remote surroundings while attending a meeting with your holographic presence among your remote colleagues. This experience, also known as telepresence, will be enabled by 6G applications belonging to the telepresence use case family [2]. The evolution and new capabilities promised by 6G, especially in the telepresence field, provide the opportunity to seamlessly merge extended reality (XR), hologram, and haptic capabilities into the day-to-day teleconference meetings.

A key enabler of 6G, artificial intelligence (AI)/machine learning (ML) will enhance the efficiency and intelligence of networks by providing a number of services to such novel 6G applications. For example, the European horizon project Hexa-X [2] mentions an AI framework that would provide AI services to various 6G applications and network functions following AI-as-a-Service (AIaaS) model.

The All-Senses meeting use case, which belongs to the telepresence family of use cases identified in 6G, will use a number of AI/ML services. These services could either be delivered by AIaaS model or through third-party, or native AI models provided by the application developers. In order to

✉ Leyli Karaçay
leyli.karacay@ericsson.com

Zakaria Laaroussi
zakaria.laaroussi@ericsson.com

Sonika ujjwal
sonika.a.ujjwal@ericsson.com

Elif Ustundag Soykan
elif.ustundag.soykan@ericsson.com

[1] Research, Ericsson, Istanbul, Turkey

[2] Research, Ericsson, Helsinki, Finland

[3] Product Security, Ericsson, Stockholm, Sweden

Published online: 05 April 2024

render realistic audiovisual and achieve co-presence, several functionalities such as stream compression and optimization, holographic rendering, and face and gaze tracking will use AI/ML [3, 4]. While commissioning additional capabilities to the applications, AI/ML expands the threat surface of the application [5]. The factors such as AI/ML model training and deployment (at edge or cloud location), additional application interfaces to consume AI services, data ownership of training data, and inherent threats to AI/ML, subject these applications to new attacks. Hence, it is necessary to understand the threat landscape in relation to the AI/ML functionality of these applications.

In our earlier work [6], we performed a generic threat modeling for the All-Senses meeting use case. We later refined our work and added AI/ML aspects to the threat modeling which resulted in [1]. In this work, we have extended our work in [1] focusing on threat modeling of AI/ML subsystem including generative large language model (we will use LLM from now on) services used in All-Senses meeting use case. We first define the All-Senses meeting use case and then illustrate how the All-Senses meeting use case utilizes numerous AI/ML services. We perform a threat analysis of this AI/ML functionality using the STRIDE threat modeling method [7] which is the most widely used method. To investigate more on the threats, we use the attack tree methodology for the selected threats. Lastly, we point out related countermeasures to defend the system against the identified threats. Our work will pave the way to understand the threat landscape of the AI/ML subsystem in the All-Senses meeting use case and shed light on LLM threats and their mitigations, thus providing a clearer picture of how to defend such systems in a holistic manner. The novelty of our work lies in the fact that we put the application of AI/ML in holographic use case perspective, how AI/ML can be utilized using native as well as pre-trained models as shown in our proposed system view, and how generic and specific AI/ML threats might affect the various stakeholders such as meeting participants and application developers.

2 Related works

The All-Senses meeting use case is different than the teleconference system as it captures holograms of remote participants to provide the sense of co-presence to the participants and provides the sense of touch and feel through haptic devices and actuators. Google's Starline project [8] provides an initial glimpse of holographic-based meetings. Although it facilitates a telepresence system between two remote participants, it lacks the haptic capabilities, thus constraining the system to be used only as a 3D meeting system. In All-Senses meetings, end devices like cameras, lidar sensors, microphones, and haptic devices capture the video, audio, and tactile motions to produce the hologram of a remote person and allow the interaction with a person's hologram. The tactile motions produce the feeling of touch and perception by triggering sensory nerve signals. The communication involved in All-Senses meeting belongs to the category of multiple sensory media (Mulsemedia) [9] and holographic-type communication (HTC) [10]. HTC stipulates requirements on network in terms of ultra-high bandwidth (of order of terabyte per second), ultra-low delay (less than 1 ms), and strict synchronization between multiple streams belonging to audio, video, and tactile motion. From network's perspective, even though this communication will be warranted by 6G capabilities, stream optimization is still a major issue in HTC. Research communities are trying to develop efficient audio and video stream compression algorithms to remove redundant streams from the transmission. From an end-device perspective, even with efficient compression techniques, HTC puts requirements in terms of storage, computation, analytics, and streaming capabilities on the end device.

In the case of constrained end devices, the solution is to offload computation (in terms of rendering, compression, and encoding) to the edge node [11]. The edge nodes can place codecs in such a way that raw content is streamed through network and can be encoded/decoded at desirable quality as per the network QoS and device capability. Looking at the 6G offerings, these services can be provided by edge network in the form of Compute as a Service (CaaS).

In the All-Senses meeting use case, we are envisioning the use of numerous AI/ML-based services. AI is also being used in generating 3D hologram in real time resulting in faster, efficient, and lightweight solution that performs well with low computational capability devices [12]. In current scenarios, a number of applications are utilizing the power of AI/ML to provide new capabilities and enhance the efficiency and accuracy of current capabilities, e.g., sentiment analysis [13], gaze correction, super-resolution, noise cancellation, and face relighting [14]. Various videoconferencing apps are already using AI-powered background noise cancellation and background obfuscation services [15]. As AI/ML techniques have shown great improvement in optimizing streaming data from multiple sources, big tech companies are using AI-based voice codecs (e.g., Google Lyra and Microsoft Satin). Due to the widespread presence of AI in telepresence and HTC, organizations are working to integrate and standardize codecs with AI [16]. Another service that could utilize AI/ML in the near future is user movement prediction which allows to optimize streaming of the holographic view based on a user's viewpoint. In addition to these services, generative AI and LLMs can act as smart meeting assistant [17] for meeting dialog transcription, creating meeting minutes, extracting action points, assigning tasks, etc. Generative AI can also be used for generating 3D holograms which can be

served as personal hologram creator to the meeting participants.

As we identify the AI/ML services that could be availed in the All-Senses meeting use case, we observe that there is no comprehensive literature available about the threat ecosystem of such futuristic applications. Raiful et al. [18] provide a generic threat analysis work on a video-conference system without taking AI/ML components into account. Similarly, in another study [19], the security assessment of meetings with end-to-end encryption in Zoom video-conference application is provided. In [20], Zoombombing which is a phenomenon where aggressors attend online meetings with the intention of disrupting the meetings and harassing its members is discussed where the first data-driven analysis of calls for Zoombombing attacks on social media is conducted. In study [21], a systematic threat analysis of extended reality systems and Metaverse is provided. They mostly emphasize on technology weaknesses, cybersecurity challenges, and users' safety concerns. In our earlier work [6], we performed a generic threat modeling for the All-Senses meeting use case. As AI/ML capabilities are being explored and utilized in a number of arenas, several organizations are working towards understanding and defining the taxonomy related to AI components (assets, asset ownership, AI lifecycle, threats) in various applications and use cases [22, 23]. The security concerns of ML have so far been the subject of numerous research articles [24–26]. In addition, Lara et al. [27] propose a methodology to identify threats in a ML-based system. Lastly, ETSI also argues that in order to secure a system, it is mandatory to secure its AI/ML components from generic and ML-specific threats [28].

3 AI/ML-enabled All-Senses meeting

All-Senses meeting allows participants to attend the meetings using their holographic presence and experience their remote surroundings through haptic capabilities. Both of these functionalities are enabled as well as facilitated by AI/ML services.

Figure 1 depicts the scenario selected for the All-Senses meeting use case. In this scenario, we have the organizer in physical location 1 giving a presentation to an audience in another physical space through the hologram. The holographic experience enables the organizer of the meeting to interact with the audience in location 2 as if she/he is in the same room. The meeting's physical location is equipped with several sensors, cameras, and microphones. In addition, the person has implanted in-body sensors and actuators. Combining inputs generated by devices and sensors helps to create the holographic experience. For instance, the data collected from location 2, such as sound and movements, allows the organizer to understand the audience's surroundings and adjust his presentation accordingly. Additionally, sensors are used to follow the movements of an organizer, ensuring that his holographic image is always in the correct position and aligned with his movements. Actuators, on the other hand, trigger nerve signals that enable sensing experience, These signals create a sensation for the organizer as she/he is present in location 2 via his hologram. Moreover, actuators allow the organizer to interact with the audience through the sense of touch despite being a hologram.

All-Senses meeting includes three different phases: pre-meeting, ongoing meeting, and post-meeting where several AI/ML services can be utilized at each phase.

- Pre-meeting phase: Services such as proposing meeting time and suggesting participants or experts for the meeting can be considered.
- Ongoing meeting phase: AI/ML can be used to provide services such as sentiment analysis, AI virtual assistant (VA), data visualization, anti-fraud systems, voice and hologram optimization, network optimization, background noise cancellation, haptic signal processing, and background blurring/object occlusion.
- Post-meeting phase: Services such as preparing meeting notes, defining and tracking action points, recommending experts, and rescheduling follow-up meetings can be provided.

In this chapter, we provide a system view of All-Senses meeting from an AI perspective. We focused on the ongoing meeting phase for the system view. Since most AI services are used in this phase, we can reflect the whole system view.

3.1 System view from AI perspective

We illustrate our system view in Fig. 2 in which we highlight the AI-ML components and services used in the All-Senses meeting. In what follows, we describe each of the key components of our system:

- ***Ecosystem:*** The ecosystem represents the physical location where the meeting is happening. In our scenario, we envision a meeting between two individuals with an audience. The medium will be through holograms equipped with sensing capabilities. Thus, the meeting location must be equipped with sensors (e.g., on-body sensors, actuators), cameras, and microphones. The on-body sensors collect vital data and the actuators trigger nerve signals which will enable sensing. We anticipate that multimedia streaming operations, such as compression, encoding, decoding, and rendering, will be handled at the edge nodes, ecosystem devices, and cloud. The edge nodes are equipped with intelligent offloading strategies, that will reduce the CPU, memory, and energy consump-

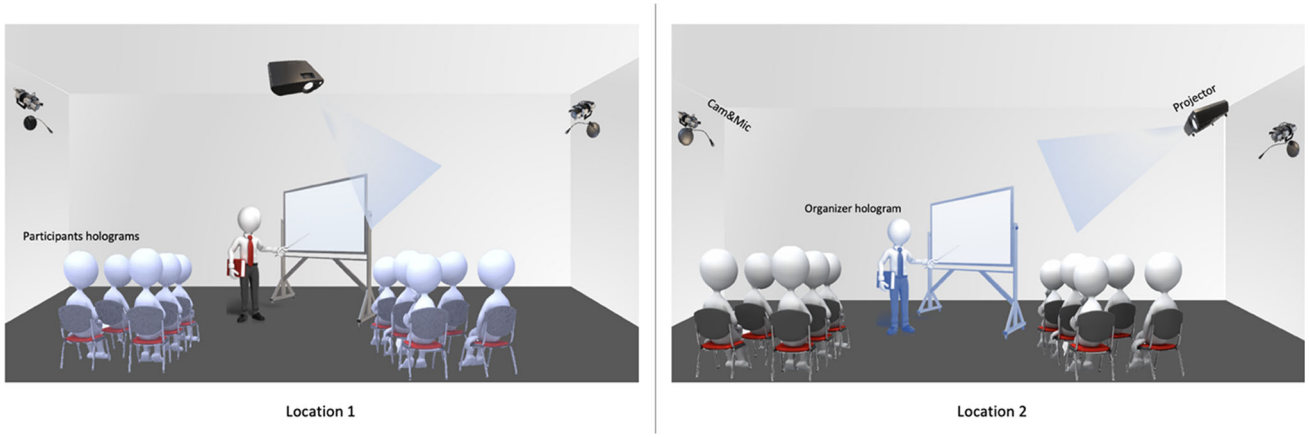

Fig. 1 All-Senses meeting selected scenario

tion of high-demanding tasks, which could otherwise put a strain on the constrained devices inside the ecosystem.

- ***Cloud:*** The collected data is maintained by the cloud to be processed and fed to the AI/ML pipeline. AI/ML pipeline refers to a series of processes such as data refinement, feature selection, model training, model optimization, and model testing. The native AI models created in this pipeline are deployed on the cloud to provide various functionalities to the holographic and multimedia communication. The AI Engine is responsible for searching and selecting the appropriate pre-trained models from the AI/ML marketplace and adapting them as per the requirement. The adaptation is done through transfer learning process. Depending on the AI model, fine-tuning methodologies can be employed. In the case of generative LLMs, fine-tuning methodologies like in-context learning can be used [29].
- ***AI/ML Marketplace:*** It is the entity that provides a plethora of pre-trained models including foundational generative LLMs like generative pre-trained transformer (GPT) tailored for a specific application or service, allowing faster and more efficient deployment.
- ***Original equipment manufacturer (OEM):*** Since there are many devices and sensors in our ecosystem, the OEM

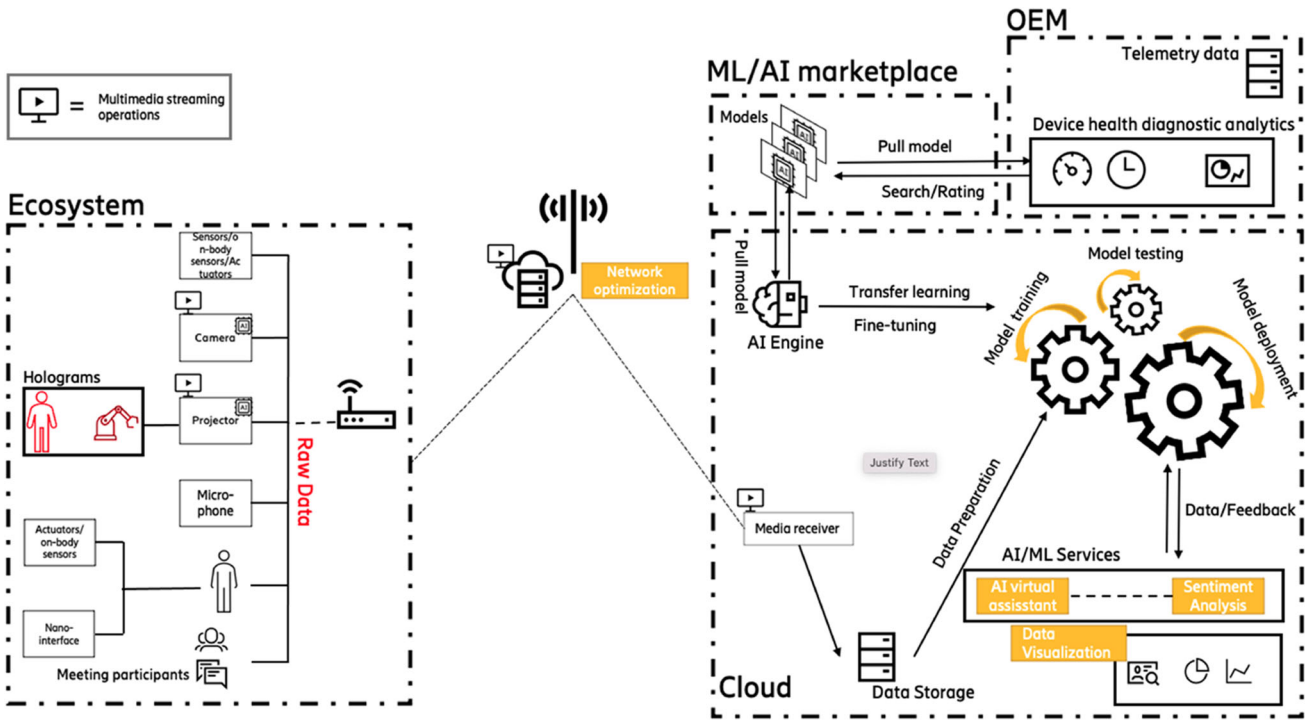

Fig. 2 Proposed system view

must push updates. The OEM might also collect some data from the ecosystem to power their device health diagnostic analytic engine in order to develop more efficient and innovative features or services for the end devices. OEM can benefit from pre-trained models in the AI/ML Marketplace to perform these analytics.

The raw data is collected from various devices within the ecosystem, and then encoding and decoding are conducted at the edge node to compress data, originating mainly from the camera, microphone, and haptic devices. After such data is obtained, it is stored in the cloud and fed to the AI/ML lifecycle to create the appropriate models. These models will be used to generate and optimize the holographic experience in terms of quality of experience (QoE). Pre-trained models may be fetched by the AI engine from the AI/ML marketplace, rather than being created from scratch. Moreover, some parts of the data can be fed to the OEM back-end servers. The servers will run prediction algorithms that can anticipate failures and then push regular updates and patches to their devices.

4 Threat modeling

Threat modeling is a technique to systematically identify threats in a system. The identified threats are later prioritized in terms of severity in order to plan mitigation steps. Many threat modeling frameworks are available such as PASTA, OWASP, OCTAVE, and STRIDE [30]. The common steps in most of the threat modeling frameworks are identifying the security objectives of the system, identifying assets that need to be protected, identifying the adversaries and their capabilities, identifying threats, and exploring possible mitigations. We decided to use STRIDE here since it is widely used, but other choices are equally possible. STRIDE stands for the six categories of threats: spoofing identity, tampering with data, repudiation, information disclosure, denial of service (DoS), and elevation of privilege (EoP). Since the original STRIDE categories do not explicitly address the AI/ML context, we define STRIDE categories in AI/ML context in order to later map an AI/ML threat to a STRIDE category. We begin threat modeling by discussing the asset taxonomy and security requirements for the assets. We define threat actors by their capabilities. We proceed by identifying the threats using STRIDE and provide the security requirement violated by the threat in a threat table. Finally, we provide guidelines for possible mitigation strategies against the identified threats.

4.1 Asset taxonomy

Our scenario in Fig. 2 outlines the interactions between the components and the assets that need to be protected. The first component is the ecosystem which includes the end devices that enable the holographic experience. These end devices belong to the Environment/Tools asset category, and meeting participants and holograms belong to the Actors asset category. The second component is the cloud that interacts with the ecosystem by collecting raw data to be processed and fed to the AI/ML pipeline; in addition, the AI engine which is responsible for fetching and selecting the needed pre-trained model from the AI/ML marketplace is counted as assets in our use case. The third component is the AI/ML marketplace which is considered in our scenario as a third-party provider of pre-trained models that facilitate the development and deployment of AI/ML services where these models are counted as assets that need to be trustworthy to be used in our system. In the fourth component, OEM, the OEM manufacturer is considered an asset that is responsible for pushing updates and patches to the end devices. As the edge node is providing computationally heavy processes, we have identified computational platforms, AI/ML platforms, haptic tools, and codec tools that belong to the Environment/Tools category of the asset. Also, processes such as encoding, decoding, synchronization, and raw data collection are identified at the edge node which belongs to the Process category of asset.

We identify assets in our system and divide them into five categories inspired by the asset taxonomy defined in the European Union Agency for Network and Information Security (ENISA) report [22], which are Actors, Data, Processes, Model, and Environment/Tools. Each category comprises related assets as depicted in Fig. 3.

4.2 Security objectives

We enumerate security objectives for the All-Senses meeting system in Table 1 and define them from an adversary's perspective. They should be interpreted as follows: "What actions of an adversary a system should deter/disallow in order to achieve a security objective." So, any violation of these security requirements is considered a threat. In addition to traditional Confidentiality, Integrity, Availability (CIA) objectives, we also define authentication, authorization, non-repudiation, robustness, safety, and privacy as additional objectives of the system. We define robustness in the context of ML models as the resilience of the ML model against adversarial attacks. In other terms, the accuracy of ML models should not change drastically in the presence of an ongoing attack. The robustness objective is important to consider here as AI/ML models are critical assets to the All-Senses meeting system and their resilience to adversarial attacks is necessary for the security and trustworthiness of the system. Adversaries aim to violate the security objectives to compromise the system. In the third column of Table 1, we present the category of threat which will be used by the adversaries to violate a security objective through a variety of methods.

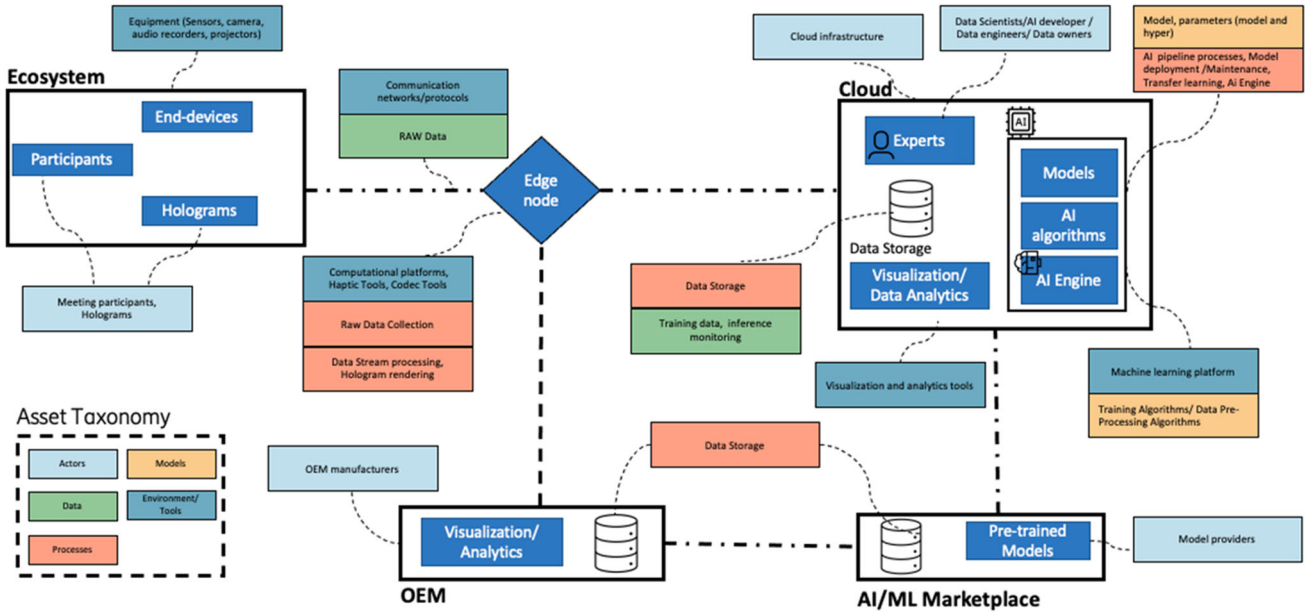

Fig. 3 Identified assets

4.3 Threat actors

Threats mostly come from two categories of sources: insiders with access authorization and outsiders without access authorization.

Insiders can be sub-categorized into authorized insiders and opportunistic insiders. A malicious meeting participant or a malicious meeting organizer can be classified under authorized insider. An opportunistic insider seizes the chance to exploit vulnerabilities or situations during virtual meetings and take use of them for his own gain.

Outsiders can be sub-categorized to unauthorized participants, hacktivists, phishers, and competitors. Unauthorized participants can be those legitimate users who are not invited to the meeting or can be non-legitimate users who attempt to join virtual meetings without proper authorization. They may monitor, intercept, and modify network traffic and gain the access rights after malicious attempts to the system. The hacktivists are driven by political or social causes, targeting virtual meetings to disrupt or spread a message during a meeting. Phishers are attempting to use links or messages to trick participants into disclosing sensitive information. And competitors are organizations or individuals seeking a competitive advantage by infiltrating and gathering information from virtual meetings.

4.4 Identified threats

In this subsection, we identify AI/ML-specific threats for All-Senses meeting use case using STRIDE. In order to apply STRIDE to ML-specific threats, some customization to STRIDE is required. Table 2 shows the proposed ML-specific definition for each STRIDE threat category. In Table 3, we map the threats associated with assets under consideration, i.e., data, processes, environment/tools, actors, and models to one or more STRIDE threat categories. In addition to STRIDE categories, we also identified threats related to safety and privacy. We add a new column in the table named as Safety/privacy which represents whether a threat is associated with privacy and/or security or neither of them. We iterate over each asset in the system and attempt to identify threats that aim to violate the security objectives. We find some generic threats applied to AI/ML and some specific holographic-related threats.

One important asset that could be targeted by adversaries in All-Senses meetings is haptic tools/platforms which enhance the overall sensory experience of users during a meeting. The threat associated with the security of this asset is haptic signal tampering where adversary can tamper with or inject malicious haptic signals while feed-backing to harm remote participants.

Another important asset that adversaries can target in next-generation meetings is the AI VA voice service which is powered by natural language processing (NLP) and ML algorithms to understand and respond to participant's commands and create a comprehensive and interactive meeting experience. The problem that comes with AI VA is that it is prone to manipulating the voice commands, which arises security and privacy concerns.

In the next section, we explain in more details about these two attacks and represent their attack trees.

Table 1 Security objectives

Security objectives	Description	STRIDE category
Confidentiality	The adversary should not be able to access any data belonging to the system, e.g., streaming raw data from participants, AI/ML specific data such as training/testing/verification data, model features, parameters and weights, and model performance metric data	Information disclosure
Integrity	The adversary should not be able to tamper, manipulate, or feed artificial or malicious data to the system The adversary should not be able to manipulate the system's process, logic, AI/ML models, etc.	Tampering
Availability	The adversary should not be able to disrupt the system or a part of the system or result in a reduced QoS of the system. The system should be able to provide services to its authorized users as per the SLA or as per the availability metric requirements	Denial of service
Non-repudiation	The adversary should not be able to non-repudiate his presence or any alteration to the system. The system should be able to trace (audit) and detect any malicious footprints/ activity in the system	Repudiation
Authentication	The adversary should not be able to gain unauthorized access to any part of the system, impersonate a legitimate/authorized user or its hologram, and use the system illegitimately	Spoofing
Authorization	The adversary should not be able to illegitimately access any component of the system. The system should ensure proper authentication and authorization check on each component/service of the system. At the same time, the adversary should not be able to move illegitimately from one part/component of a system to another	Elevation of privilege
Robustness	The adversary should not be able to drastically reduce the accuracy and performance of AI/ML models. In addition, the AI/ML models should be resilient against adversarial (evasion) attacks	Tampering
Privacy	The actors of the system (including adversary) should not be able to collect private data from users (meeting participants) of the system without their consent. They should also not use the collected user data for any other purpose without informing and taking consent from users Also, the system should provide assurance that users will not be uniquely identified from the data collected by the system	Information disclosure
Safety	The actors of the system (including adversary) should not be able to tamper with the system in a way to harm the meeting participants or to physically destroy/vandalize the system assets. The system should provide some assurance related to the physical security of participants and against unwarranted and inappropriate haptic touches to the participants	Not applicable

5 Attack trees

Attack trees are defined as logical graphs used to visually and systematically illustrate different attack paths that an adversary may use to compromise the security of the system [31]. Attack trees consist of nodes and edges, where each node represents points of compromise or attack vectors and the edge represents the causal relationship between the nodes. The root of the attack tree represents the adversary's objective or end goal which can be achieved by taking any of the multiple paths starting from any leaf node to the root node. In this way, attack trees provide a hierarchical representation of a sequence of actions from an adversary to achieve an objective. Through a threat modeling process, security experts can benefit from attack trees as a powerful tool for proactive cybersecurity to enhance an organization's security

Table 2 ML-specific definition of STRIDE threats

STRIDE threat	ML-specific definition
Spoofing	Involves threats where the adversary tries to gain unauthorized access to AI/ML system or deceive the system by masquerading identities of a legitimate entity
Tampering	Involves threats where the adversary generates or alters the data used throughout a model's lifecycle to compromise model's integrity
Repudiation	Adversary performs malicious activities and denies that the model's output has been produced by it in any way
Information disclosure	Involves threats where an adversary learns and leaks data, models, and other information from AI/ML system
Denial of service (DoS)	Involves threats where an adversary tries to overwhelm the AI/ML system with malicious or adversarial inputs to degrade performance or exploit the resource requirements of ML models to exhaust system resources
Elevation of privilege (EoP)	Involves threats where an adversary tries to gain higher access rights to AI/ML system than intended to perform unauthorized actions

posture and foresee potential threats. In this vein, we will discuss the generated attack trees of two selected threats from Table 3.

Figure 4 illustrates the attack tree for the haptic signal tampering threat in the All-Senses meeting use case. In this case, the threat is related to tampering with haptic signals in a meeting involving holographic devices with AI/ML capabilities. Haptic signals provide a more immersive experience in holographic environments. The use of AI/ML for haptic signal processing could potentially broaden the attack surface, making the system more vulnerable to attacks; for instance, an outsider adversary can disrupt the confidentiality of the system by gaining unauthorized access to AI/ML model parameters and training data related to haptic signals by having illegitimate access to the infrastructure (e.g., equipment). Conversely, jamming attacks interfere with network infrastructure and device signals' availability through physical or electromagnetic interference. This attack will disrupt normal functioning, challenging accessing or launching the holographic experience by degrading service quality for All-Senses meeting participants. From an integrity perspective, an outsider or insider adversary may manipulate or distort the haptic feedback generated by AI/ML models by altering the tactile haptic signals, causing confusion and misinformation, or even compromising the safety of users. An outsider or insider adversary can achieve this by modifying frequencies or injecting false haptic signals.

Figure 5 illustrates the attack tree for disruption of AI VA service in the All-Senses meeting use case which targets the AI VA asset. AI/ML can be integrated into VA to respond to user input and improve the services provided. As a result, it may increase the threat surface of the system and cause it to be more vulnerable to attacks. For instance, an outsider adversary may disrupt the confidentiality of the system by impersonating an authorized user's voice by misleading VA's voice recognition model, and manipulate authorized user's setting in VA which can lead to disruption in the intended functionality. From an integrity aspect, an outsider adversary can add carefully crafted noise to input stream to degrade the performance of AI model in VA, as well as exploit the vulnerabilities in NLP system to manipulate the semantic of the user's command to VA system. Additionally, from an availability perspective, DoS attacks which aim to overwhelm the AI model in VA can be conducted by submitting computationally costly queries or using ultrasonic waves to create commands which are not audible to human ears but can disrupt the VA service. Lastly, intended/unintended voice command can also be submitted to VA system which may lead to unauthorized/unintended actions or responses.

In order to tackle the threats posed by each of the attacks, the next section offers countermeasure mechanisms to mitigate them.

Table 3 Threat table

Asset type	Asset name	Threat	STRIDE mapping	Safety/privacy	Description
Data	Raw Data	Data poisoning/tampering	Tampering/DoS	-	Adversary can tamper with the data collected from ecosystem e.g. adding superficial data, adversarial data to distort the ML model during training
Data	Stored Data	Label manipulation	Tampering, DoS	-	Adversary can change the labels of data (random or specific data) to be used for training the ML models in order to mislead the ML model
Data	Stored Data	Data tampering	Tampering	-	Adversary can manipulate the data stored in the cloud
Data	Stored Data	User data leakage	Information disclosure	Privacy	Adversary can access private user data such as audio files, authentication credentials, biometric data stored in cloud if access to data is not secured
Data	Pre-trained Models	Data leakage	Information disclosure	Privacy	LMs may leak sensitive data in the response accidentally or via crafted prompts as they are not able to hide or filter out sensitive information. This may lead privacy and even intellectual property issues if the training data is collected via web-scraping
Processes	AI VA	Disruption of AI VA services	Spoofing, tampering, DoS	Privacy	Adversary can disrupt the services provided by VA by impersonating a legitimate user or tampering the data, as well as conducting DoS attacks or injecting intended/unintended voice commands
Processes	Raw Data Collection	Participants behavioral tracking	Information disclosure	Privacy	Adversary can infer private attribute related to the target participants or if it has prior knowledge about the target, inspects the target's habit/behavior
Processes	Raw Data Collection	Unauthorized data collection	Information disclosure	Privacy	Adversary including dishonest OEMs, service providers, and cloud infrastructure providers can collect more data from participants than they are authorized to, without participants consent and they may even sell that data to third-parties
Processes	Model Training	ML training process tampering	Tampering, DoS	-	Adversary can tamper with model training process in order to behavior e.g. selecting poisoned dataset for training, stopping the optimization cycles prematurely resulting in an inaccurate model
Processes	Model Adaption/Transfer Learning	Backdoor insertion attack	Tampering	-	Adversary can insert a backdoor in the ML model during the transfer learning process that can be triggered at the inference phase
Processes	AI Engine	Malicious model selection	Tampering	-	Adversary can manipulate the AI engine logic to select malicious models from AI/ML marketplace
Processes	AI Engine Transfer Learning	Malicious code insertion attack	Tampering	-	Adversary might embed malware or ransomware (EvilModel) in the ML model which can be extracted during the transfer learning process to evade detection from antivirus engines
Environment/ Tools	Analytic Tools	Analytic tool tampering	Tampering	-	The analytic tool's input can be altered by an adversary to give inaccurate results

Table 3 continued

Asset type	Asset name	Threat	STRIDE mapping	Safety/privacy	Description
Environment/Tools	Communication Protocols	Eavesdropping attack	Information disclosure	Privacy	Adversary can eavesdrop/sniff the communication between meeting participants e.g. in case insecure communication protocols are used between end devices and edge node. Adversaries could also employ machine learning techniques to analyze audio and video streams and infer sensitive information about users or organizations by exploiting patterns in network traffic
Environment/Tools	Communication Networks	DoS	DoS	-	Adversary can cause DoS attack in communication network resulting in loss of quality of service of meeting. It would also hinder the remote assistance/learning capability due to poor streaming quality of holograms of objects
Environment/Tools	Haptic Tools/Platforms	Haptic signal tampering	Tampering	Safety	Adversary can tamper with or inject malicious haptic signals and feedback to harm remote participants or to vandalize remote objects
Environment/Tools	End devices/Sensors Infrastructure	Physical attacks/physical damage	DoS, Tampering	-	Adversary in the same physical area may damage the end devices making them unavailable, or they may inject malware in the devices to sniff network communication
Environment/ Tools	End devices/Sensors Infrastructure	DoS attacks	DoS	-	Adversary can perform a denial-of-service attack on the end devices making hem unavailable or resulting in loss of QoS of meeting
Environment/ Tools	End devices/ Sensors Infrastructure	Ransom attacks	Information disclosure	Privacy	Adversaries may gain access to a microphone, haptic devices, etc. and record participants' behavior and interactions in the meeting. Later, they may threaten to release these recordings publicly unless the user pays a ransom
Environment/ Tools	Third-party/ ML platform	Supply chain/ attacks	Tampering	-	Adversaries can inject malicious code or tamper with code in ML platform libraries in order to affect the ML models generated using that ML platform
Actors	Meeting Participants	Harassment of participants	-	Privacy/ Safety	Participants can be harassed by other remote participants in the form of unwarranted and inappropriate touch or physical harm
Actors	Meeting Participants	Phishing attacks	Tampering, Elevation of privilege	-	AI algorithms may be used to generate convincing phishing messages targeting meeting participants. In this way, the adversary can gain access to systems, networks, or data without the appropriate permissions. The attack may introduce malware which can corrupt or alter data within a system
Actors	Meeting Participants	Deepfake attacks	Spoofing	Privacy	Adversaries may use dee pfake technology powered by AI to impersonate meeting participants and gain unauthorized access to systems and sensitive data

Table 3 continued

Asset type	Asset name	Threat	STRIDE mapping	Safety/privacy	Description
Actors	Meeting Participants	Overreliance to LLMs	Tampering	-	LLMs may confidently produce inaccurate responses by their generative nature. Meeting Participants could trust the output without verifying which may lead to wrong decisions or actions
Actors	Hologram	Digital assault	Tampering	Safety	Adversary can alter the hologram in a way to make participants act in an unexpected way which may lead to physical harm
Actors	Hologram	Perception Manipulation Attacks	Tampering, Spoofing	Safety	Adversary could manipulate the human multi-sensory perceptions of the physical world to influence user's decision-making, e.g., overlay a virtual part of the robot with another part, making the user interact incorrectly with the robot or slowing their reactions
Models	Model Parameters	Model stealing/IPR theft	Information disclosure	-	Adversary can steal proprietary model parameters or the complete ML model, e.g., insecure access to ML model can lead to model stealing due to unlimited inferences by adversary
Models	Model Parameters	Model Manipulation	Tampering, DoS	-	Adversary can manipulate model parameters to alter AI system's behavior which could result in model evasion or misclassification, etc.
Models	Model	Poisoned pre-trained model from marketplace	Tampering, Spoofing	-	Adversary can place malicious/adversarial pre-trained model on ML marketplace which could be selected by AI engine
Models	Model	Adversarial inference attack	Tampering, DoS	-	Adversary can tamper the input data to the ML model in order to evade/deceive the ML model at inference phase. E.g., tracking the facial recognition model in order to bypass authentication
Models	Model	Direct prompt injection	Tampering, Information disclosure	Privacy	Privacy & Adversary can modify the system level prompt restrictions to jailbreak the LLMs and override prompt level controls. This threat occurs when adversary directly interacts with the LLMs
Models	Model	Indirect prompt injection	Tampering, Information disclosure	Privacy	Adversary sends prompts by interacting with a service using the LLM program or uploading a file as an input to LLM and tries to access the back-end service using the LLM as a proxy. This may cause unauthorized access to the back-end service

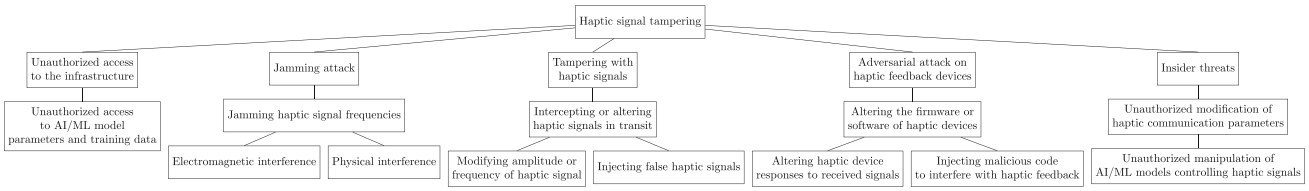

Fig. 4 Attack tree: haptic signal tampering threat

6 Mitigation strategies

In this section, we discuss some countermeasures for the threats identified in the threat Table 3. It should be noted that these measures should be supported by related cybersecurity controls and implemented based on the severity of the risk that should be evaluated after threat analysis by performing an impact and likelihood analysis. Most of the attacks can be mitigated by implementing fundamental security and monitoring capabilities throughout the AI/ML development and deployment lifecycle [32]. However, for AI/ML-specific attacks, additional mitigation strategies are needed. The brief summary of the mitigation strategies with respect to asset types is illustrated in Table 4.

To protect Data asset, data pre-processing and cleaning are important to remove misleading data in the training dataset. In case data is acquired from third-party data owners, it should be properly sanitized before injecting it into the AI/ML data pipeline. Monitoring the performance of the machine learning system regularly involves analyzing model outputs, evaluating model behavior on new data, and conducting periodic reviews of the training dataset which is beneficial to identify poisoned data. Also, data protection for data at rest, in transfer, and in use is inevitable using data encryption, integrity, data signing, and data isolation techniques. To protect the privacy of participants, privacy should be integrated into data protection measures using a privacy-by-design approach [33]. Finally, access control policies play a vital role in protecting data by ensuring that only authorized individuals or entities have appropriate access rights. When integrating LLMs into the systems, it is crucial to know that sensitive information may be unintentionally disclosed by LLMs in their responses, which could result in unauthorized data access and privacy violations. So, meeting participants' or in general user inputs should not be used to enhance LLM training data. LLM applications should carry out sufficient data sanitization to prevent user data from getting into the training data. Additionally, if a foundational LLM is used or a third-party LLM service is onboarded, it is important to check terms and conditions of the model so that users can choose not to have their data used in the training model and are informed about how their data is treated.

To protect Process asset, maintaining the integrity of the processes involved in the AI/ML pipeline is crucial, e.g., transfer learning and pre-trained model employment processes should be protected from any malicious activities. Robust access control policies based on minimum privilege access principle should be applied on the basis of the explicit role of an actor. Logging and monitoring are crucial to detect any unauthorized behavior in the processes including model training processes. In addition, to detect model training anomalies, model validation and monitoring techniques should be implemented. The All-Senses meeting use case collects more data from the ecosystem compared to traditional video-conference applications; hence, privacy by design is a crucial requirement. As participants' data related to sentiments, behavior, eye movement, fingerprint, facial data, etc. could be collected by application providers to deliver an immersive experience, informing participants about the user data collection and requiring explicit consent, complying with regulations such as GDPR and data handling agreements between parties should be a high priority.

To protect Environment/Tools asset, using third-party components should be taken seriously so as to minimize the risk of supply-chain attacks, be it a third-party ML platform, haptic tools and platform, and the cloud infrastructure. Therefore, using secure platforms, tools, and libraries and implementing application security is also important

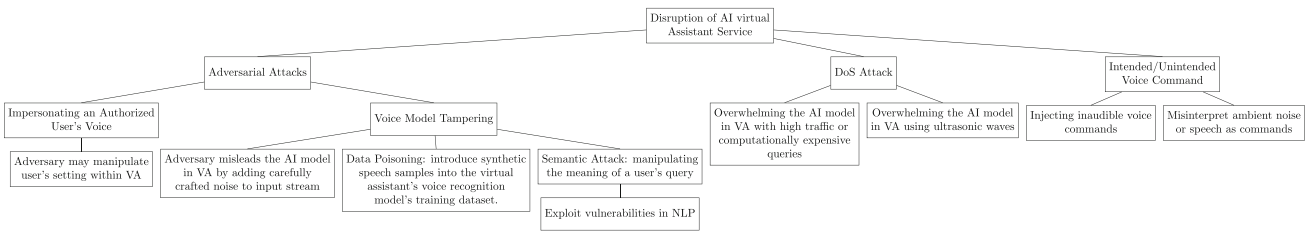

Fig. 5 Attack tree: disruption of AI VA service threat

Table 4 Summary of mitigation strategies

Asset type	Mitigation strategy
Data	Training data sanitization and pre-processing, model validation and monitoring, access control policies
	data protection, privacy by design
Process	Process integrity, access control policies, logging and monitoring, model validation and monitoring
	privacy by design
Environment/Tools	Secure communication protocols, access control policies, logging and monitoring, physical protection of
	haptic tools/platform/end devices, secure platforms, tools, and libraries, application security
Actors	Physical protection of projector, software-assisted safety measures, enhanced privacy for the users
Model	Adversarial training, input data refinement, model regularization, model validation and monitoring
	confidential computing, AI/ML marketplace security, AI/ML API security (output minimization
	and obfuscation)

for securing AI/ML since software and AI/ML operations are interacting. Unauthorized access to microphone, haptic devices, etc. should be prevented with access control mechanisms. Physical access to ecosystem premise should be secured to deter the end-device physical attacks. Logging and monitoring of the environment and the tools are also inevitable to detect any unexpected movement or situation. Secure communication protocols such as Transport Layer Security (TLS) protocol need to be used between end devices and edge node to encrypt data, authenticate data origin, and ensure message integrity to prevent adversary to eavesdrop/sniff the communication.

To protect Actor asset, providing safety and privacy is vital. To make the participants safe from digital assault, virtual harassment, and perception manipulation attacks, developers should include software-assisted safety measures [34] in the application. Safety measures should also be supported by enhanced user privacy techniques using a privacy-by-design approach as mentioned earlier. Furthermore, end devices like projectors should be protected from unauthorized physical access. Generative LLMs may create a false sense of trust in their responses, confidently answering any question without regard for the accuracy of the response. This hallucination effect should be considered when the responses are used, e.g., responses should not be directly used in decision-making without cross-checking. Additionally, fine-tuning or parameter-tuning methods can be used to improve the accuracy.

To protect Model asset, different model hardening methodologies such as adversarial training or input data refinement for evasion attacks, training data distribution monitoring [35] and training data sanitization [36] against data poisoning attacks, can be used. Note that the mitigation strategies may change depending on the model algorithm, and more detailed information on different approaches can be found in [37]. Model regularization [38] mitigates the impact of individual instances, including poisoned ones, by reducing overfitting and preventing excessive weight allocation. Model validity framework should be in place to assess model's inference accuracy and request the retraining of the models if accuracy and validity drop from an acceptable level. Confidential computing techniques can be utilized to prohibit malicious access to proprietary training algorithms and model parameters to protect the intellectual property rights [39]. AI/ML marketplace should be properly protected from unauthorized parties and monitored for any malicious behavior, e.g., unauthorized upload/download of pre-trained models. Since AI/ML API is the entry point for internal as well as external party/services, it is essential to incorporate API security[40] in the AI/ML API, e.g., robust access control on API end point. Also, output minimization and obfuscation techniques can be employed to respond to API requests in order to prevent inference attacks.

When it comes to protecting LLMs, all above-mentioned methodologies are applied and needed as well. However, specific attacks to LLMs, direct and indirect prompt injection attacks, require special attention since LLMs have the potential to transform a wide range of use cases. These attacks enable adversaries to override original instructions and employ controls aiming to manipulate how LLM works in the system they are integrated. By crafting the prompt carefully, an adversary can attempt to influence the behavior of the model and make it generate responses that are biased,

harmful, or deceptive. Input validation and sanitization of prompts are the most important prevention methods to protect models against prompt injection attacks. Additionally, continuous monitoring and auditing of model outputs can help detect and mitigate any potential issues as the adversaries will keep exploring new bypass methods. More detailed scenarios for LLMs can be found in [41]

7 Conclusion

With the widespread evolution of AI/ML technologies in 6G networks, numerous new use cases are made possible. Our proactive approach towards addressing the potential threats to 6G not only reaffirms our dedication to staying ahead of the curve but also aligns seamlessly with the evolving technologies such as AI/ML in 6G networks. In this work, we examined the All-Senses meeting, a 6G use case, and evaluated the AI/ML services it offers as well as potential security risks. It is vital to comprehend the threats of the use case that makes use of AI/ML capabilities and look into potential risks arising from the architecture of the AI/ML pipeline as well as the intrinsic characteristics of the AI/ML components in the use case.

We determined the potential AI/ML services that are used at the various stages of the All-Senses meeting use case. With the aim of identifying and analyzing the threats, we also derived the security objectives of the system, ecosystem actors, assets, and their interactions. We note that, in comparison to previous work, new features for this use case, such as the ability to manipulate human multi-sensory perceptions of the physical world, tamper with or inject malicious haptic signals and feedback, holographic and haptic capabilities aided by AI/ML technologies, and integration of generative AI and LLMs can open up new entry points and new threats. We extended our analysis for the selected threats by using attack tree methodology with graphical representation to understand how attacks might succeed. Our evaluations show that there are opportunities for additional research and study that involve exploring these new aspects and enhancing security measures for critical threats. For future works, a risk assessment study can be performed to identify the potential impact and likelihood for the threats which will help in prioritizing mitigation efforts and allocating resources effectively.

Author contribution All authors contributed equally to the publication and reviewed the manuscript.

Funding This work was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) through the 1515 Frontier Research and Development Laboratories Support Program under Project 5169902, and has been partly funded by the European Commission through the Horizon Europe/JU SNS project Hexa-X-II (Grant Agreement no. 101095759).

Data availability No datasets were generated or analyzed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Karaçay, L, Laaroussi Z, Ujjwal S, Soykan EU (2023) On the security of 6G use cases: AI/ML-specific threat modeling of All-Senses meeting. In: 2023 2nd International conference on 6g networking (6GNet), pp 1–8. https://doi.org/10.1109/6GNet58894.2023.10317758
2. Khorsandi BM (2022) Targets and requirements for 6G - initial E2E architecture. report, European Union's Horizon 2020 research and innovation programme. https://hexa-x.eu/wp-content/uploads/2022/03/Hexa-X_D1.3.pdf
3. Kononenko D, Lempitsky V (2015) Learning to look up: realtime monocular gaze correction using machine learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4667–4675
4. Lombardi S, Simon T, Saragih J, Schwartz G, Lehrmann A, Sheikh Y (2019) Neural volumes: learning dynamic renderable volumes from images. ACM Trans Graph 38(4). https://doi.org/10.1145/3306346.3323020
5. Mauri L, Damiani E (2022) Modeling threats to AI-ML systems using STRIDE. Sensors 22(17):6662
6. Laaroussi Z, Ustundag Soykan E, Liljenstam M, Gülen U, Karaçay L, Tomur E (2022) On the security of 6G use cases: threat analysis of 'All-Senses Meeting'.In: 2022 IEEE 19th annual consumer communications & networking conference (CCNC), pp 1–6. https://doi.org/10.1109/CCNC49033.2022.9700673
7. Shevchenko N, Chick TA, O'Riordan P, Scanlon TP, Woody C (2018) Threat modeling: a summary of available methods. Technical report, Carnegie Mellon University Software Engineering Institute Pittsburgh United
8. Lawrence J, Goldman DB, Achar S, Blascovich GM, Desloge JG, Fortes T, Gomez EM, Häberling S, Hoppe H, Huibers A, Knaus C, Kuschak B, Martin-Brualla R, Nover H, Russell AI, Seitz SM, Tong K (2021) Project Starline: a high-fidelity telepresence system. ACM Trans Graph (Proc SIGGRAPH Asia) 40(6)
9. Akyildiz IF, Guo H (2022) Holographic-type communication: a new challenge for the next decade
10. Clemm A, Vega MT, Ravuri HK, Wauters T, Turck FD (2020) Toward truly immersive holographic-type communication: challenges and solutions. IEEE Commun Mag 58(1):93–99. https://doi.org/10.1109/MCOM.001.1900272
11. You D, Doan TV, Torre R, Mehrabi M, Kropp A, Nguyen V, Salah H, Nguyen GT, Fitzek FHP (2019) Fog computing as an enabler for immersive media: service scenarios and research opportunities. IEEE Access 7:65797–65810. https://doi.org/10.1109/ACCESS.2019.2917291
12. Shi L, Li B, Kim C, Kellnhofer P (2021) Author Correction: Towards real-time photorealistic 3D holography with deep neural networks. Nature 593(7858):13–13. https://doi.org/10.1038/s41586-021-03476-5
13. Patel K, Mehta D, Mistry C, Gupta R, Tanwar S, Kumar N, Alazab M (2020) Facial sentiment analysis using AI techniques: state-of-the-art, taxonomies, and challenges. IEEE Access 8:90495–90519. https://doi.org/10.1109/ACCESS.2020.2993803

14. NVIDIA (2020) NVIDIA announces cloud-AI video-streaming platform to better connect millions working and studying remotely
15. Martinek R, Kelnar M, Vanus J, Koudelka P, Bilik P, Koziorek J, Zidek J (2015) Adaptive noise suppression in voice communication using a neuro-fuzzy inference system. In: 2015 38th International conference on telecommunications and signal processing (TSP), pp 382–386. https://doi.org/10.1109/TSP.2015.7296288
16. MPAI Community. http://mpai.community
17. Bibhudatta, D (2023) Generative AI will transform virtual meetings
18. Hasan R, Hasan R (2021) Towards a threat model and security analysis of video conferencing systems. In: 2021 IEEE 18th annual consumer communications & networking conference (CCNC), pp 1–4. IEEE
19. Isobe T, Ito R (2021) Security analysis of end-to-end encryption for zoom meetings. IEEE Access 9:90677–90689
20. Ling C, Balcı U, Blackburn J, Stringhini G (2021) A first look at Zoombombing. In: 2021 IEEE symposium on security and privacy (SP), pp 1452–1467. https://doi.org/10.1109/SP40001.2021.00061
21. Qamar S, Anwar Z, Afzal M (2023) A systematic threat analysis and defense strategies for the metaverse and extended reality systems. Comput Secur 103127
22. Challenges AC (2020) AI cybersecurity challenges,threat landscape for artificial intelligence. report, European Union Agency for Cybersecurity, ENISA. https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges
23. Tabassi E, Burns KJ, Hadjimichael M, Molina-Markham AD, Sexton JT (2019) A taxonomy and terminology of adversarial machine learning. NIST IR, 1–29
24. Papernot N (2018) A marauder's map of security and privacy in machine learning: an overview of current and future research directions for making machine learning secure and private. In: Proceedings of the 11th ACM workshop on artificial intelligence and security, pp 1–1
25. Papernot N, McDaniel P, Sinha A, Wellman MP (2018) SoK: security and privacy in machine learning. In: 2018 IEEE european symposium on security and privacy (EuroS & P), pp 399–414. IEEE
26. Barreno M, Nelson B, Sears R, Joseph AD, Tygar JD (2006) Can machine learning be secure? In: Proceedings of the 2006 ACM symposium on information, computer and communications security, pp 16–25
27. Mauri L, Damiani E (2021) STRIDE-AI: an approach to identifying vulnerabilities of machine learning assets. In: 2021 IEEE international conference on cyber security and resilience (CSR), pp 147–154. IEEE
28. Industry Specification Group (ISG) Securing artificial intelligence (SAI) (2021) Securing Artificial Intelligence (SAI). ETSI, France
29. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JDea (2020) Language models are few-shot learners. In: Larochelle, H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) Advances in neural information processing systems, vol 33, pp 1877–1901. Curran Associates, Inc., ???. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf
30. Selin J (2019) Evaluation of threat modeling methodologies
31. National Cyber Security Center (2023) Using attack trees to understand cyber security risk. https://www.ncsc.gov.uk/collection/risk-management/using-attack-trees-to-understand-cyber-security-risk
32. Microsoft (2021) Microsoft AI Security Risk Assessment, Best Practices and Guidance to Secure AI Systems
33. Casella D, Lawson L (2022) AI and privacy: everything you need to know about trust and technology. https://www.ericsson.com/en/blog/2022/8/ai-and-privacy-everything-you-need-to-know
34. Boyd C (2022) Meta blows safety bubble around users after reports of sexual harassment
35. Hong S, Chandrasekaran V, Kaya Y, Dumitras T, Papernot N (2020) On the effectiveness of mitigating data poisoning attacks with gradient shaping. ArXiv:2002.11497
36. Nelson B, Barreno M, Chi FJ, Joseph AD, Rubinstein BIP, Saini U, Sutton C, Tygar JD, Xia K (2008) Exploiting machine learning to subvert your spam filter. In: USENIX workshop on large-scale exploits and emergent threats
37. Oprea A, Vassilev A (2023) Adversarial machine learning: a taxonomy and terminology of attacks and mitigations (draft). Technical report, National Institute of Standards and Technology
38. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT press, ???
39. Ustundag Soykan E, Karaçay L, Karakoç F, Tomur E (2022) A survey and guideline on privacy enhancing technologies for collaborative machine learning. IEEE Access 10:97495–97519. https://doi.org/10.1109/ACCESS.2022.3204037
40. Foundation TO (2023) OWASP API Security Project.https://owasp.org/www-project-api-security/
41. Foundation TO (2023) OWASP top 10 for large language model applications.https://owasp.org/www-project-top-10-for-large-language-model-applications