(IJIASE) 2024, Vol. No. 10, Special Issue **e-ISSN: 2454-9258, p-ISSN: 2454-809X**

Assessing the Impact of AI and Virtual Reality on Strengthening Cybersecurity Resilience Through Data Techniques

Praveen Kumar Maroju

QA Lead Architect
Clientserver Technology Solutions
SAN ANTONIO Texas USA

[1]*Received: 30 May 2024; Accepted: 20 August 2024; Published: 23 August 2024*

ABSTRACT

Technological developments in AI present civilization with enormous prospects. At the same time, there is a growing need to address new implications. Because of this, the emphasis is frequently placed on moral and secure design to prevent inadvertent mistakes. On the other hand, methods focused on cybersecurity and AI safety also take into account instances of deliberate evil, such as immoral and hostile AI design. Recently, there has been a similar focus on malevolent actors in relation to virtual reality (VR) security and safety. Thus, even while the nexus of AI and VR (AIVR) presents a plethora of advantageous opportunities for cross-fertilization, considering the possible socio-psycho-technological ramifications, it is imperative to foresee future malevolent AIVR design from the outset. This study examines the potential application of Generative AI (deepfake techniques) for deception in immersive journalism, as a simplified example. From an immersive co-creation perspective, we believe that defenses against such future AIVR safety hazards associated to lying in immersive contexts should be envisaged trans disciplinarily. We initially drive a cybersecurity-focused process to produce defenses through immersive design fictions.

INTRODUCTION

To ensure that the benefits of advancements in AI are realized, it's crucial to address potential risks associated with AI development and deployment from the outset. As a result, AI safety and ethics are increasingly recognized as vital components of research across various subfields and internationally. Typically, AI safety and ethics focus on implementing ethical and secure AI systems while avoiding design and operational errors. However, from a cybersecurity perspective, it is also important to consider the threat posed by malicious and unethical actors. These adversaries may deliberately attack AI systems or design them with harmful intentions.

In the context of virtual reality (VR), recent research on VR security and safety, as well as mixed reality, aligns with this cybersecurity-focused approach. It emphasizes the need to anticipate potential misuses and attacks by malicious entities. Malicious design represents a worst-case scenario in AI safety, where the system is controlled by the attacker, granting them extensive adversarial capabilities with minimal restrictions. The same concerns apply to malicious design in VR.

Given the promising potential of synergies between AI and VR technologies, it is crucial to proactively identify possible misuses of these integrations. Recent recommendations highlight the importance of considering malicious use cases early on for secure development across various mixed reality applications. This paper examines the intersection of AI and VR (AIVR) with a focus on unethical design practices that we term "immersive falsehood." Immersive falsehood involves deliberately crafted synthetic immersive environments intended for malicious purposes.

[1] *How to cite the article:* Maroju P.K..; Assessing the Impact of AI and Virtual Reality on Strengthening Cybersecurity Resilience Through Data Techniques; *International Journal of Innovations in Applied Sciences and Engineering;* Special Issue 2024, Vol 10, No. 1, 1-9

Fig 1: Cyber Resilience privilege

To illustrate, we analyze a speculative but plausible scenario: targeted disinformation through VR news content enhanced by Generative AI, such as future extensions of VR deepfakes. As immersive journalism—news formats that offer participatory, first-person experiences of recreated events—advances, the creation of VR news productions is becoming a reality. While this convergence of AI, VR, and immersive journalism presents innovative opportunities, it also amplifies the risk of malicious AIVR designs and disinformation.

Although advanced AIVR applications are still in their infancy and may currently be considered niche, analyzing these potential risks is already valuable for the separate fields of AI and VR. Insights from worst-case scenarios can inform defenses against simpler, related threats. For instance, understanding how to counteract misleading immersive journalism in VR can help in addressing deceptive non-immersive disinformation videos created with Generative AI, which affect fewer sensory modalities.

Moreover, these insights can enhance security awareness among VR users by highlighting potential adversarial tactics and manipulation techniques. Confronting malevolent AIVR designs prompts researchers and society to address the challenge of developing defenses against immersive falsehoods. One approach to this challenge is using design fictions—narratives or prototypes in text, audio, video, or VR formats—that project future technological scenarios and inform security strategies.

MALEVOLENT ACTORS AND FALSEHOOD IN AIVR

Malevolent creativity refers to the intentional use of creativity to achieve harmful objectives. In AI, as in cybersecurity, this form of creativity can lead to ongoing conflicts between adversaries and defenders. Just as dynamic interactions between defenders and ethical hackers, as well as a focus on safety, contribute to a more balanced security landscape in cybersecurity, similar approaches in AI can enhance security. This includes efforts in adversarial machine learning, where a growing body of research addresses both adversarial attacks and adaptive defense mechanisms.

Currently, malevolent actors can exploit various vulnerabilities in AI systems, leading to issues such as adversarial examples, poisoning attacks, machine learning backdoors, and model theft. Specific instances of malicious AI design include the development of harmful AI software, misuse of automated drones or autonomous vehicles, and the creation of Generative AI for purposes like disinformation, extortion, and defamation.

In virtual reality (VR), malevolent actors can potentially inflict psychological or physical harm. This can occur through various means, such as displaying or overlaying offensive content, harassing users in social virtual spaces, manipulating users' physical movements to harmful locations, or inducing dizziness and confusion. In extreme cases, manipulating subtle elements like the frequency of visual stimuli could pose risks to neurologically vulnerable individuals.

Additionally, privacy threats in social VR settings may arise, such as identity theft of user avatars and unethical tracking of private channels. An emerging concern is the unethical creation and dissemination of synthetic non-consensual VR content, which could be worsened by creating virtual replicas or modifications of real individuals without consent. Future developments in VR deepfake technology, which are already technically feasible, could exacerbate these issues.

AI-enhanced malevolent VR design could be used for large-scale manipulative purposes, including immersive disinformation. While VR-based immersive journalism offers unprecedented access to news through sights, sounds, and emotions, this feature could be exploited for deception, especially when combined with Generative AI. As noted in a recent article, the fusion of deepfakes and VR could undermine trust in shared information and lead to highly manipulated content across various channels.

A. Malicious Design of Generative AI

The most advanced forms of Generative AI currently available to malevolent actors are deepfake techniques utilizing deep learning (DL) tools. Although commonly associated with face-swapping, deepfakes extend beyond this, encompassing alterations in speech, text, body movements, and images across various domains. While deepfakes offer many positive and innovative applications in fields like gaming, entertainment, healthcare, education, and privacy-preserving journalism (see e.g. for an overview), their potential for misuse could undermine these benefits.

Malicious uses of deepfakes include disinformation, revenge, extortion, sabotage, defamation, fraud, tools for other cybercrimes, scams, impersonation, obfuscation, tampering with legal evidence, and physical harm. In the following section, we will outline some practical risk scenarios to illustrate these concerns.

The following high-level processes can be exploited across various domains for malicious Generative AI design: 1) replacement, 2) reenactment, 3) image synthesis, 4) speech synthesis, 5) synthetic text generation, 6) adversarial perturbation, and 7) automated disconcertion.

Process 1, commonly associated with facial replacement or face-swapping in computer vision, has been infamously used in a public defamation video that falsely portrayed journalist Rana Ayyub in explicit contexts.

Process 2 involves facial reenactment, where the facial features of one individual are transferred to another's face, allowing attackers to impersonate and control the target’s speech and actions. This technique poses increasing risks for audiovisual journalism content.

Process 3 enables the creation of fake artifacts that appear to be portraits of real individuals. This has been used to generate misleading profile pictures on social media to create fake personas, and for disinformation and espionage. For instance, deepfakes have been applied to medical imagery to falsify diagnostic features, as demonstrated with lung cancer scans.

Process 4 includes voice cloning, where deep learning techniques are used to impersonate individuals. An example is a UK CEO impersonation scam, where an employee was tricked into transferring a significant sum of money.

Process 5 involves synthetic text generation, such as using fine-tuned versions of the Generative Pre-trained Transformer (GPT-2) model to create text that mimics political disinformation.

In process 6, adding adversarial perturbations to deepfake material aims to evade detection by deepfake detectors—a technique known as “adversarial deepfakes”. This strategy can be used to obscure other cybercrimes or conceal inauthentic content related to disinformation campaigns. Such tactics could have severe forensic implications and negatively impact the information ecosystem. They may also complicate content filtering for sensitive issues like

terrorist propaganda or child abuse, especially if illegal authentic material is modified with deepfakes for identity obfuscation and then perturbed to bypass detection.

Process 7, automated disconcertion, refers to the confusion and diminished credibility of audio, visual, and textual evidence resulting from the widespread availability and misuse of these deepfake techniques. In forensic contexts, this can manifest as the "liar's dividend," where genuine evidence is undermined by the sheer volume of manipulated content. On a societal and interpersonal level, this makes it increasingly difficult to resolve suspicions of falsehood, providing a strategic advantage to malicious actors engaged in targeted disinformation. For example, the recent failed military coup in Gabon was partly fueled by the belief that an official presidential video was a manipulative deepfake.

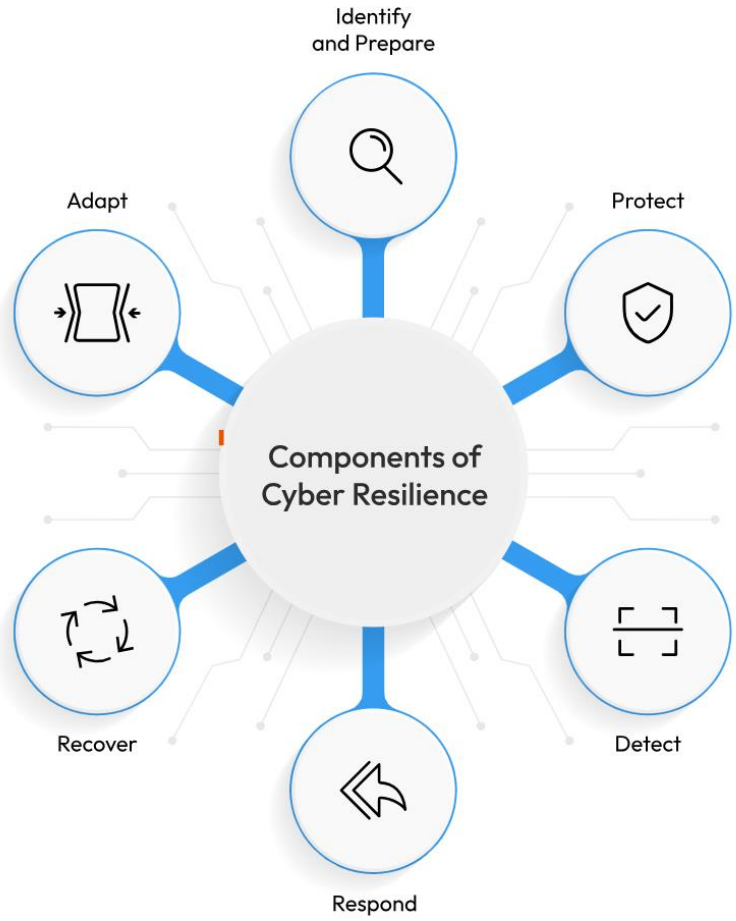

Fig 2: Components of Cyber Resilience

B. Immersive Journalism, VR and Disinformation

A notable vision for the emerging field of immersive journalism (IJ), as articulated by De la Peña—often referred to as the “godmother of VR”—is to rekindle “the audience’s emotional involvement in current events,” addressing a perceived indifference to human suffering. IJ is believed to enhance empathy and evoke a sense of awe and wonder. Recent studies suggest that IJ can also lead to positive changes in attitudes. Initially, it was proposed that VR news content based on 3D video would provide a more realistic experience compared to conventional formats, a concept that remains relevant with today's highly detailed 3D reconstructions.

Furthermore, VR has the potential to transform news reporting by allowing users to be virtually present at news events, offering a new level of immersion beyond traditional photographic and acoustic documentation. VR news experiences are associated with higher telepresence and greater credibility compared to standard news formats. The use of VR

headsets in IJ could enable unique experiences of immersive 3D "spatial journalism", introducing user-directed spatial dynamics that enhance the sense of presence.

Despite these promising developments and the existence of various IJ formats, including AR frameworks, 360-degree reports, and drone-based immersive news, IJ is still in its early stages. Currently, the most common forms are 360-degree videos and mobile VR settings, partly due to the complexity and expense of VR content creation.

Over the past two decades, several early immersive journalism (IJ) formats in VR have been developed. One of the first VR news stories from The New York Times (NYT) was "The Displaced", which, while available only as a 360-degree film through the NYT app (and sometimes debated as not strictly VR content), allowed viewers to visually explore the impact of war on three displaced children from different countries.

Another notable example is "Project Syria," which offered an immersive VR experience of a bomb explosion in Aleppo and a refugee camp, viewable with Oculus Rift or HTC Vive. Similarly, "Assent" was a VR documentary available for Oculus Rift that depicted military executions in Chile from the perspective of the filmmaker's father.

Additionally, a VR piece designed to raise awareness about the conditions at Guantánamo Bay prison involved reconstructing the prison for Second Life and later Unity3D. These examples of VR journalism provide a unique perspective by immersing viewers in a space where they perceive credible actions as happening in real-time, involving their own bodies in the experience.

According to De la Peña, it is this blend of presence, experience plausibility, and bodily engagement that offers a "profoundly different way to experience and ultimately understand the news," providing insights that would be impossible without such immersion.

The very attributes that make immersive journalism (IJ) in VR compelling also render it attractive to malicious actors. The ability of VR to create presence, immersion, empathy, and a sense of credibility could be exploited to promote falsehoods propagated by manipulative entities . Various IJ formats could thus be misused for propaganda and disinformation.

To clarify, instead of the broader term "fake news," which overlaps with both disinformation and misinformation, we use the term "disinformation" as defined by the UK government. Disinformation refers to the deliberate creation and dissemination of false or manipulated information intended to deceive and mislead audiences, often for harmful, political, personal, or financial purposes.

In a 2017 interview with Quartz, De la Peña acknowledged the potential for VR to be misused for propaganda and poor journalism, emphasizing that the issue lies more with the creators than with the medium itself. Thus, the primary concern is the malevolent creativity of malicious creators, although unintentional errors are also a factor. Similar to the dynamics in cybersecurity and deepfakes, an arms race between attackers and defenders is likely to emerge in the realm of disinformation. Future developments in IJ and VR could potentially follow this trend.

C. Manipulated VR News and False Memory Construction

As VR technologies become more affordable and widespread, the risk of immersive falsehoods created by malicious actors in immersive journalism (IJ) settings grows. Sanchez has described potential dystopian scenarios where users could be engulfed in a world of fake news. Similarly, Uskali and Ikonen highlight the need for IJ experts to be vigilant about sophisticated manipulation and disinformation operations, noting that "our brain believes so strongly in what it sees in VR that distinguishing fake news from real news might become challenging".

A significant concern for the future of IJ is the deliberate creation of false memories through experiential VR news content. Studies show that immersive VR experiences, compared to traditional desktop displays, can leave a more lasting impression by integrating visual, vestibular, and proprioceptive senses [. This immersive quality could provide a new avenue for malicious actors to exploit VR journalism for disinformation purposes. Liv and Greenbaum argue that false memories can be generated through various means, including VR, to promote fake news on both individual and mass scales. Research by Frenda et al. indicates that false memories can be used for political manipulation, with greater success if the content aligns with existing preferences. Additionally, a study found that children are particularly

susceptible to false memories in VR, and a 2018 study demonstrated that even minimal exposure to misleading images and narratives can lead to false memories in adults. Given these findings, it is highly plausible that hyper-realistic IJ experiences in VR could amplify such psychological effects.

Building on the analysis of creating enduring false memories for disinformation, we identify three key processes that could facilitate this goal: 1) persuasive spatial dynamics engineering, 2) memory-centered sensory stimulation, and 3) information gathering.

Process 1 involves systematically designed techniques that enhance spatial awareness, perception, and orientation in VR environments. This could include methods like integrating 3D minimaps to create a more immersive and convincing experience.

Process 2 focuses on utilizing specific sensory stimuli to improve memory consolidation. For instance, future adversaries could leverage olfactory displays for VR—currently in development but not yet commercially available. Given that olfactory pathways are distinct and olfactory memories are particularly vivid and resistant to forgetting and interference, such displays could significantly enhance the impact of VR experiences. These olfactory elements could be integrated into VR headsets or used as standalone devices.

Process 3 encompasses techniques like social engineering and open-source intelligence gathering , which involve collecting publicly available data on individuals to understand their pre-existing preferences and beliefs. This information could be used to tailor VR content more effectively, ensuring that the disinformation aligns with the target's personal context and increases its effectiveness.

CYBERSECURITY-ORIENTED IMMERSIVE DEFENSES

In the previous Section II, we examined the range of possibilities available to malevolent actors within AI and VR technologies, specifically highlighting their potential use in disinformation within immersive journalism. In this section, we will outline a cybersecurity-focused approach for developing defense strategies against adversaries operating at the intersection of AI and VR (AIVR).

To address this, we begin by discussing the importance of threat modeling, a critical component in both cybersecurity and recent security research on machine learning. Threat modeling involves clearly defining the goals, capabilities, and knowledge of potential adversaries. Before delving into the development of generic defense measures for immersive AIVR, we first present a threat model tailored to our specific use case of malicious AIVR design, providing a foundational illustration for our defense strategy.

A. Threat Modelling for Malevolent AIVR Design Use Case

• Adversarial goals: In our use case, the adversary's objective is to engage in targeted disinformation by integrating AI and VR tools within immersive journalism (IJ) settings. The adversary aims to manipulate the opinions, attitudes, and views of specific IJ audiences according to a carefully crafted scheme. Specifically, the goal is to alter a source set of conceptions \(S \) to a target set of conceptions \(T \) within a particular context. These sets may vary in both content and the level of confidence assigned to each element. The adversary seeks to deceive and mislead for political, personal, or financial gain, or simply to cause harm. Ultimately, this goal represents a form of microtargeted disinformation in IJ.

• Adversarial knowledge: We assume that all Generative AI components involved are accessible in white-box settings, and the same transparency applies to the VR content creation used in IJ experiences. The adversary can obtain publicly available information about the victims and may also use social engineering to acquire additional personal data. Malicious Generative AI, including deepfakes and VR deepfakes, can be seen as a form of adversarial examples targeting humans—designed to deceive individuals by manipulating their preferences, beliefs, and perceptions through a carefully crafted sensory experience. When the adversary has successfully gathered crucial personal information about the victims, this can be considered a grey-box setting, representing a middle ground between black-box and white-box levels of adversarial knowledge.

• Adversarial capabilities: In terms of Generative AI, the adversary can utilize the seven processes identified earlier: replacement, reenactment, image synthesis, speech synthesis, synthetic text generation, adversarial perturbation, and automated disconcertion. Regarding VR content creation, the adversary has full control over the design and integration of multimodal materials such as images, videos, and audio samples. They can also apply the three relevant processes for VR content creation: persuasive spatial dynamics engineering, memory-centered sensory stimulation, and information gathering.

Overall, the adversary can employ a total of ten different processes to achieve microtargeted disinformation. It is important to note that in practice, the range of possible techniques could be broader and is limited only by the adversary's creativity. Therefore, defenses should be viewed as incremental measures rather than definitive solutions.

B. Immersive Design Fictions for AIVR Safety

Design fiction (DF) provides a valuable framework for "HCI and design researchers to co-create, explore, and speculate about the future". Recently, Houde et al. demonstrated the effectiveness of co-creation DF in addressing near-term AI safety concerns related to Generative AI use cases. Given this success, we propose that DF is an effective methodology for developing defenses against potential AIVR safety risks.

To ensure a systematic approach, we recommend anchoring future AIVR DF efforts in threat models. Additionally, the principle of requisite variety in cybernetics, which states that "only variety can destroy variety", implies that to develop effective defenses against adversaries operating across physical, virtual, and immersive realms, an immersive perspective is essential. This immersive approach is crucial for creating meaningful solutions to any malicious AIVR design involving immersive falsehoods.

Interestingly, VR has been suggested as a powerful platform for DF due to its "higher level of immersion and sense of embodiment". Thus, AIVR safety can benefit from insights gained through AIVR DF, just as DF can advance understanding in fields like cybersecurity, social psychology, affective science, law, and journalism.

Given our threat model, it's evident that design fictions for malicious AIVR use cases must address a complex socio-psycho-technological landscape encompassing immersive, digital, and physical elements with significant cognitive and emotional implications. Addressing this complexity necessitates a transdisciplinary approach.

Design fictions should focus not only on proactive defenses but also on reactive mechanisms [8]. Proactive defenses aim to prevent malevolent actors from distributing harmful VR content in the first place. This could involve preventative measures implemented before content deployment, potentially utilizing tools similar to those used for deepfake detection. However, due to the fallibility of human knowledge, the limitations of AI detection systems, and the unpredictability of human creativity, it is crucial to develop reactive defense measures. These would come into play after users have been exposed to manipulated VR news content.

It’s important to clarify that we do not view design fiction (DF) as a tool for predicting the future. The unpredictable nature of future knowledge means that extrapolations are constrained by current understanding, and we must remain prepared for unforeseen developments. DF cannot anticipate the consequences of knowledge that has yet to be created. However, it can generate plausible counterfactual scenarios that may prove invaluable.

Organizationally, we envision a preparation phase before the DF process begins, involving the creation of an immersive prototype. This prototype could be a simple immersive multimodal narrative incorporating audiovisual, olfactory, or tactile elements (e.g., as seen in recent MIT deepfake storytelling projects). Ideally, this would be a VR prototype.

We identify three distinct groups for this process:

1. Creators of the immersive prototype

2. Designers with expertise in AIVR

3. A multidisciplinary team with knowledge spanning various technological areas related to AI and VR

The suggested sequence for the immersive DF process is illustrative and not prescriptive:

1. Designer Co-Creation Session: AI and VR designers develop two threat models:

- Threat Model 1: Represents a malicious AIVR design scenario that is technically feasible today.

- Threat Model 2: Envisions a scenario that could become feasible within the next five years based on current knowledge.

2. Participant Introduction to AIVR: Designers provide a high-level overview of the current state-of-the-art in AI and VR technologies to a multidisciplinary audience.

3. Designer Narrative: Designers present Threat Model 1 to the audience.

4. Participant Co-Creation Session: Inspired by Threat Model 1, participants create a new Threat Model 3, reflecting what they believe might be technically feasible in five years based on their current knowledge.

5. Participant Narrative: Participants present Threat Model 3 to the designers.

6. Narrative Comparison: Designers present Threat Model 2, and participants compare it to Threat Model 3.

7. Immersive Session: Both designers and participants experience a short immersive prototype that illustrates Threat Model 0 (pre-fabricated by the prototype creators). This session includes:

- An immersive journalism piece (ideally in VR) featuring a real but unknown event.

- A piece featuring disinformation inspired by the Threat Model from Subsection III-A.

Users are not informed which piece is real and which is manipulated until the end of the experience.

8. Common Defense Co-Creation Session: Designers, participants, and makers collaborate to develop proactive and reactive defenses against Threat Models 0 to 3. They also explore potential adaptive attacks in response to the proposed defenses.

CONCLUSION

Recent research into the safety and security of AI and VR highlights the need to go beyond traditional approaches in designing ethical and safe systems by anticipating intentional exploits from unethical and malicious actors. To address this, we conducted a proactive cybersecurity-focused analysis of malicious designs at the intersection of AI and VR, or AIVR. Given that this field is still emerging, it is crucial to establish robust defenses from the outset rather than retrospectively. For instance, we applied our analysis to immersive journalism, illustrating how malevolent actors could leverage Generative AI and VR to create (microtargeted) disinformation and immersive falsehoods. This use case underscores the need for both proactive and reactive immersive defenses to address the socio-psycho-technological impacts of such malicious activities.

To develop effective defense measures, we introduced a cybersecurity-oriented approach to immersive co-creation design fictions, ideally conducted in VR. In essence, immersive AIVR co-creations can enhance AIVR safety. Although such co-creations may not offer a complete solution to counter malicious designs, they are valuable for incremental and on-demand updates for various AIVR safety scenarios. Furthermore, immersive design fictions informed by security practices offer a promising method to use VR as a rich counterfactual experiential testbed, extending to scenarios involving unethical actors.

Recent futures exercises have highlighted AI-generated fake content as a significant potential threat for AI-enabled crime. Generative AI technologies, like deepfakes, could be misused to create false memories. Considering VR's ability to facilitate long-lasting memories, these AIVR synergies pose risks if exploited by malicious actors. Future research could explore the psychological effects of false memories induced through such exploits and consider the potential of immersive cognitive-affective debiasing measures using AIVR technology.

REFERENCES

[1] A. Dafoe, "AI governance: a research agenda," *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK*, 2018.

[2] V. Dignum, "AI is multidisciplinary," *AI Matters*, vol. 5, no. 4, pp. 18–21, 2020.

[3] D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. Dragan, "Cooperative inverse reinforcement learning," in *Advances in neural information processing systems*, 2016, pp. 3909–3917.

[4] J. Leike, D. Krueger, T. Everitt, M. Martic, V. Maini, and S. Legg, "Scalable agent alignment via reward modeling: a research direction," *arXiv preprint arXiv:1811.07871*, 2018.

[5] D. Peters, K. Vold, D. Robinson, and R. A. Calvo, "Responsible AI – Two Frameworks for Ethical Design Practice," *IEEE Transactions on Technology and Society*, vol. 1, no. 1, pp. 34–47, 2020.

[6] N. Soares and B. Fallenstein, "Agent foundations for aligning machine intelligence with human interests: a technical research agenda," in *The Technological Singularity*. Springer, 2017, pp. 103–125.

[7] A. Asilomar, "Principles.(2017)," in *Principles developed in conjunction with the 2017 Asilomar conference [Benevolent AI 2017]*, 2018. Authorized licensed use limited to: Indian Institute Of Technology (IIT) Mandi. Downloaded on August 02,2024 at 05:06:08 UTC from IEEE Xplore. Restrictions apply.

[8] N.-M. Aliman, P. Elands, W. H¨urst, L. Kester, K. R. Th´orisson, P. Werkhoven, R. Yampolskiy, and S. Ziesche, "Error-Correction for AI Safety," in *International Conference on Artificial General Intelligence*. Springer, 2020, pp. 12–22.

[9] M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, B. Filar *et al.*, "The Malicious Use of Artificial Intelligence: Forecasting," *Prevention, and Mitigation*, 2018.

[10] F. Pistono and R. V. Yampolskiy, "Unethical Research: How to Create a Malevolent Artificial Intelligence," *CoRR*, vol. abs/1605.02817, 2016. [Online]. Available: http://arxiv.org/abs/1605.02817

[11] R. V. Yampolskiy and M. Spellchecker, "Artificial intelligence safety and cybersecurity: A timeline of AI failures," *arXiv preprint arXiv:1610.07997*, 2016.

[12] K. Pearlman, "Virtual Reality Brings Real Risks: Are We Ready?" *USENIX Association*, 2020.

[13] P. Casey, I. Baggili, and A. Yarramreddy, "Immersive virtual reality attacks and the human joystick," *IEEE Transactions on Dependable and Secure Computing*, 2019.

[14] A. Gulhane, A. Vyas, R. Mitra, R. Oruche, G. Hoefer, S. Valluripally, P. Calyam, and K. A. Hoque, "Security, Privacy and Safety Risk Assessment for Virtual Reality Learning Environment Applications," in *2019 16th IEEE Annual Consumer Communications & Networking Conference (CCNC)*. IEEE, 2019, pp. 1–9.

[15] S. Baldassi, T. Kohno, F. Roesner, and M. Tian, "Challenges and new directions in augmented reality, computer security, and neuroscience–part 1: Risks to sensation and perception," *arXiv preprint arXiv:1806.10557*, 2018.

[16] J. A. De Guzman, K. Thilakarathna, and A. Seneviratne, "Security and privacy approaches in mixed reality: A literature survey," *ACM Computing Surveys (CSUR)*, vol. 52, no. 6, pp. 1–37, 2019.