Malicious Design in AIVR, Falsehood and Cybersecurity-oriented Immersive Defenses

Nadisha-Marie Aliman
Department of Information and Computing Sciences
Utrecht University
Utrecht, Netherlands
n.aliman@uu.nl

Leon Kester
Intelligent Autonomous Systems
TNO Netherlands
The Hague, Netherlands
leon.kester@tno.nl

***Abstract*—Advancements in the AI field unfold tremendous opportunities for society. Simultaneously, it becomes increasingly important to address emerging ramifications. Thereby, the focus is often set on *ethical* and safe design forestalling unintentional failures. However, cybersecurity-oriented approaches to AI safety *additionally* consider instantiations of intentional malice – including *unethical* malevolent AI design. Recently, an analogous emphasis on malicious actors has been expressed regarding security and safety for virtual reality (VR). In this vein, while the intersection of AI and VR (AIVR) offers a wide array of beneficial cross-fertilization possibilities, it is responsible to anticipate future malicious AIVR design from the onset on given the potential socio-psycho-technological impacts. For a simplified illustration, this paper analyzes the conceivable use case of Generative AI (here deepfake techniques) utilized for disinformation in immersive journalism. In our view, defenses against such future AIVR safety risks related to falsehood in immersive settings should be transdisciplinarily conceived from an immersive co-creation stance. As a first step, we motivate a cybersecurity-oriented procedure to generate defenses via immersive design fictions. Overall, there may be no panacea but updatable transdisciplinary tools including AIVR itself could be used to incrementally defend against malicious actors in AIVR.**

***Index Terms*—AI Safety, VR, AI, Immersive Journalism, Disinformation, HCI, Design Fiction, Psychology, Cybersecurity**

I. Motivation

For humans to benefit from progresses in the AI field, it is essential to early start to also tackle the potential risks associated with AI development and deployment. In the light of the foregoing, AI safety and AI ethics considerations are gradually being recognized as indispensable component of AI research efforts across multiple research subfields [1]–[6] at an international level [7]. Commonly, methods in AI safety and AI ethics focus on how to implement ethical and safe AI systems and how to avoid unintentional failure modes related to design-time mistakes and operational failures. However, from a cybersecurity-oriented view in AI safety [8]–[11], it has been emphasized to *additionally* consider the existence of malicious and *unethical* actors. Such adversaries can intentionally launch malicious attacks on deployed AI systems or themselves craft AI systems with *intentional malice in design*. Concerning VR settings, recent work on security and safety for VR [12]–[14] and also more generally mixed reality [15]–[18] is in line with this cybersecurity-oriented AI safety perspective and stresses the need to anticipate misuses and attacks by malicious entities. Generally, the AI risks embodied by malicious design can be understood as worst-case scenarios in AI safety given that the system is owned by the attacker allowing maximal adversarial capabilities with minimal restrictions in white-box settings[1] [19]. Obviously, the same holds analogously for malicious design in VR.

Hence, given the promising avenues that beneficial synergies between AI and VR technologies started to bring forth [20], it seems important to proactively identify possible misuses of such synergies. In fact, the early consideration of individual use cases involving malevolent actors has been recently explicitly recommended for a security-aware development across diverse mixed reality instantiations [17]. In this paper, we focus on malevolent design at the intersection of AI and VR (AIVR). More specifically, we zero in on intentionally performed unethical AIVR design that construes what we call *immersive falsehood*. We regard immersive falsehood as a landscape of deliberately designed synthetic immersive realities for malicious purposes. For a graspable analysis and by way of illustration, we use the not yet prevalent but cogitable use case of targeted disinformation via VR news contents potentiated by Generative AI (such as e.g. via exploits using future extensions of VR deepfakes [21]). Indeed, regarding the information ecosystem in the near future, progresses in the nascent field of immersive journalism (which refers to news formats allowing participatory first-person experiences of recreated news events and situations [22]) already include the creation of VR news productions [23]. Thereby, while the convergence of AI, VR and such experiential news could provide a unique window of opportunity for innovations, it could also simultaneously exacerbate the space of possibilities for *malicious AIVR design* and disinformation.

Note that even if advanced AIVR applications might currently belong to a niche in its infancy, the analysis of this particular type of risk could be already useful *today* for AI, VR and their safety separately. The reason being that instructive insights gained from such worst-case scenario considerations might already be applicable to simpler cases. For instance, when being equipped with methods on how to defend against

[1]Simply put, in security, white-box settings refer to cases where the adversary has full knowledge about the internal implementation of the system.

978-1-7281-7463-1/20/$31.00 ©2020 IEEE
DOI 10.1109/AIVR50618.2020.00031

misleading immersive journalism experiences in VR, one might be better informed on how to tackle the consumption of manipulative disinformation videos crafted with Generative AI which are non-immersive and affect much less sensory modalities. Concurrently, a similar set of methods might be useful to sensitize VR users and raise security awareness on the procedure of potential adversaries including their manipulation techniques. Finally, the mere confrontation with instantiations of malevolent AIVR design leads involved researchers (and ultimately society) to face the relevant issue on how to possibly generate *defense methods* against vividly experienced immersive falsehood. One possibility to address this task is the use of *design fictions* [24]–[27] known from human computer interaction (HCI) and which have been recently also applied to near-term and long-term security issues of modern technologies including Generative AI [28]. (Co-creation design fictions can be used for technological future projections by experts in the form of e.g. narratives or construed prototypes that can be represented in text, audio or video formats but also in VR environments [29].) The next Section II collates technical and psychological information on malicious design in the context of Generative AI and immersive journalism. Then, Section III first discusses parameters relevant to a cybersecurity-oriented modelling of adversarial capabilities and goals. Subsequently, we elaborate on how to – on this basis – co-create defenses using immersive design fictions. Thereafter, we conclude in Section IV and provide incentives for future work.

II. Malevolent Actors and Falsehood in AIVR

Malevolent creativity [30] can be described as the deliberate utilization of creativity in the service of harmful goals. As in cybersecurity, malevolent creativity applied to AI may fuel incessant races between adversaries and defenders. However, as in cybersecurity, dynamic exchanges between defenders and ethical hackers and the practice of considering safety aspects, attacks and defenses can contribute to a more informed and balanced security ecosystem [10]. Early analogous efforts can be already observed in the field of adversarial machine learning where an increasing number of publications on both adversarial AI attacks and adaptive defense methods are produced [31]. Already in current AI contexts, it is technically feasible for malevolent attackers to for instance intentionally cause a variety of failures ranging from exploitation of vulnerabilities against adversarial examples over poisoning attacks and machine learning backdoors to model thefts [32]. Regarding malevolent AI design [10], [19] itself, feasible examples include among others AI-based malicious software [33], misuse of automated drones [8], [34] or autonomous vehicles [35] and *malicious design of Generative AI* [28] for disinformation, extortion and defamation.

In VR, it is in principle practicable for malevolent actors to cause psychological or physical harm [18] e.g. by displaying or overlaying offensive undesirable contents [13], enacting harassment in social virtual spaces [36], by controlling the physical movements of the user towards maliciously chosen physical locations or by deliberately inducing dizziness and confusion [13]. In extreme cases, physical harm could be caused for instance by manipulating subtle elements such as the frequency of visual stimuli threatening hereto neurologically vulnerable individuals [15]. Malevolent actors could also threaten privacy in social VR settings [37] e.g. via identity thefts of user avatars [18] linked to the unethical tracking of multiple private channels. Furthermore, an already emerging phenomenon is the unethical crafting and sharing of synthetic non-consensual VR contents [38] – which could be exacerbated with perceived virtual replica or tailored modifications of existing humans [39] if performed non-consensually. Corresponding endeavors could be fueled by future extensions of VR deepfake methods that are technically already feasible [21], [38]. Ultimately, it is conceivable that AI-aided malevolent VR design could also be utilized e.g. for manipulative purposes at a larger scale taking the form of immersive disinformation [18]. In fact, while one advantage of the already existing VR-based immersive journalism [23] is the *"unprecedented access to the sights and sounds, and possibly feelings and emotions that accompany the news"* [22], this feature could be systematically exploited in order to deceive – especially when amplified with Generative AI. Incisively, a recent online article expounded that combining deepfakes and VR may *"damage the trust in shared information"* and could lead to *"extremely manipulated content across various channels"* [40].

A. Malicious Design of Generative AI

The currently most sophisticated instances of Generative AI that are potentially available to malevolent actors are the so-called "deepfake" techniques harnessing deep learning (DL) tools. While often associated with face-swapping methods, the range of deepfake applications transcends those contexts and comprises not only modifications of faces in image and video artifacts but also extends to speech, text and body motions as well as images in other domains. Thereby, it is important to note that deepfakes simultaneously open up a variety of beneficial and forward-looking applications (see e.g. [41] for an overview) in areas such as gaming, entertainment, health care, education or even privacy-preserving journalism. Here, we are concerned with potential misuses which if ignored, could also compromise or overshadow the unfolding of positive impacts of these technologies. Potential harmful and malicious adversarial goals to design deepfakes comprise among others disinformation, revenge, extortion, sabotage, smearing, frauds, crafting a tool for other cybercrimes, scams, impersonation, obfuscation, tempering with legal evidence and physical harm [42]–[44]. In the next paragraph, we introduce a set of practically relevant risk instantiations for illustrative purposes.

The following exemplary high-level processes could be instrumentalized across different domains for malicious Generative AI design: 1) *replacement*, 2) *reenactment*, 3) *image synthesis*, 4) *speech synthesis*, 5) *synthetic text generation*, 6) *adversarial perturbation* and interestingly 7) *automated disconcertion*. The most popular application for *process 1* is

certainly facial replacement (aka face-swapping) in the computer vision domain. Such a DL-based facial replacement has been for instance used for a public defamation video shared across ca. 40000 individuals portraying the journalist Rana Ayyub in pornographic contexts she never partook. Concerning *process 2* which often involves a type of puppetry via facial reenactment where facial features of a driving source entity are transferred to the face of a target, they *"give attackers the ability to impersonate an identity, controlling what he or she says or does"* [44]. With increasing generative capabilities, it is easily conceivable that it could become more and more problematic for audiovisual journalistic contents. Moreover, *process 3* facilitates the generation of fake artifacts perceived as portraits of possibly existing individuals. It has been already utilized to generate misleading profile pictures on social media to simulate fake personas [45] and has been harnessed for disinformation [46] and even espionage attempts [47]. Another example of malicious DL-based image synthesis is the generation of deepfakes in the domain of medical imagery to add or remove diagnostic features for which a proof-of-concept has been recently implemented as applied to scans for lung cancer [48]. *Process 4* has been for instance utilized for a type of DL-based voice cloning facilitating an impersonation of the CEO of a company in the UK where an employee could be convinced to transfer a significant amount of money [49]. Very recently, *process 5* instantiated in the form of DL-based natural language generation with a fine-tuned version of the known Generative Pre-trained Transformer (GPT-2) model has been argued to be able to formulate textual messages resembling political disinformation [50].

When it comes to *process 6*, the key motivation of adding adversarial perturbations to a previously crafted material to evade deepfake detectors (a technically feasible strategy denoted "adversarial deepfakes" [51]) could be to disguise other cybercrimes or to conceal inauthentic contents related to disinformation campaigns [45]. Its future real-world instantiations could lead to severe forensic consequences [44] and could have nefarious impacts on the information ecosystem. Beyond that, it could also lead to repercussions regarding content filters related to terroristic propaganda or child abuse [34] (which is for instance conceivable if illegal authentic material is first modified via deepfakes for identity obfuscation [52] and subsequently adversarially perturbed [51] to evade deepfake detectors). Last but not least, an interim retrospective view of this short non-exhaustive enumeration of processes that can be exploited for malevolent Generative AI design reveals the need to consider the socio-psychological and forensic impacts *of their mere existence*. In fact, with *process 7* of automated disconcertion, we refer to the automatically eventuated mechanism that is brought forth by the very availability of these processes which are potentiated by the malicious Generative AI design itself. In forensics, it materializes in the form of the liar's dividend [44] seemingly taking away the general credibility of audio, visual and textual samples. At the societal and interpersonal but also intrapersonal level, it means that founded or unfounded suspicions of falsehood might turn out to become harder and harder to resolve in practice. Needless to say that the rather diffuse automated disconcertion can represent a strategical advantage for malicious actors interested in forms of targeted disinformation. In fact, a recent failed military coup in the context of pre-existing political unrest in Gabon was partially grounded in the proliferation of the wrong assumption that an official presidential video represented a manipulative deepfake video [53]–[55].

B. Immersive Journalism, VR and Disinformation

One striking vision for the nascent field of immersive journalism (IJ) as revealed by De la Peña (who has also been called the "godmother of VR" [56]) was the explicit goal to reinstitute *"the audience's emotional involvement in current events"* [22] which seemed to exhibit a certain degree of indifference towards human suffering. It has been argued that IJ can promote empathy [57] as well as a sense of awe and wonder [58]. Moreover, a recent study found that it can foster postive attitudinal changes [59]. Further, it was initially postulated that VR news contents when *"based on 3D video rather than on 3D synthetic modeling and animation"* [22] would offer an even more realistic framing than conventional formats. This may also apply to VR content creation with modern highly detailed and realistic 3D reconstructions [20], [60], [61]. Beyond that, VR may offer *"a powerful plattform to re-create news events, taking the idea of photographic documentation of reality or acoustical recordings to an entirely new level in which the user can be virtually present at a news event and experience it as a witness or even as a participant"* [23]. Interestingly, VR news experiences have been associated with higher telepresence and even elevated news credibility [62] when compared with standard news consumption forms without VR exposure. IJ experienced with VR headsets could allow unique experiences of immersive 3D "spatial journalism" [58] via *"the introduction of user-directed spatial dynamics, adding a new level of presence"* [63]. Despite these promising avenues and the fact that there exist multiple types of IJ including AR frameworks, 360-degree reports [64] and drone-based immersive news [58], IJ is still in its infancy and the most widespread pieces correspond to either 360-degree video productions or mobile VR settings [23], [65] which is also linked to the fact that VR content creation is still relatively complex and expensive nowadays [66].

Nevertheless, multiple early IJ formats in VR have been developped in the last two decades. The first VR news story of the New York Times (NYT) (albeit only as 360-degree film downloadable from a NYT app which some would strictly speaking not label as VR content [67]) termed "The Displaced" [23] was focused on three children from different nations displaced by war and allowed a visual exploration of the effects of the devastation. Furthermore, "Project Syria" facilitated an immersive VR experience featuring a bomb explosion in Aleppo and a refugee camp [23] that could be viewed with Oculus Rift or HTC Vive while "Assent" was devised as a VR documentary that could be viewed with Oculus Rift depicting the witnessing of military executions in

Chile from the perspective of the maker's father. Another IJ piece in VR that was made available to the public was crafted to raise awareness concerning the detention conditions at the Guantánamo Bay prison and was based on a re-construction of this prison for Second Life and later for Unity3D. In these examples of VR journalism, a unique grasp of the situation becomes possible by *"transferring people's sensation of place to a space where a credible action is taking place that they perceive as really happening, and where, most importantly, it is their very body involved in this action"* [22]. In a nutshell, according to De la Peña, it is this combination of presence, the plausibility of the experience and the embodied active sampling of the environment that facilitates this *"profoundly different way to experience the news, and therefore ultimately to understand it in a way that is otherwise impossible, without really being there"* [22].

However, this set of attributes of IJ in VR make it at the same time lucrative for malicious actors. It is easily conceivable that such unique immersive experiences can also create presence, immersion, empathy and a sense of credibility in the context of falsehood advocated by manipulative entities [68]. Different IJ formats could accordingly be misused for propaganda and disinformation. For clarity, instead of using the broader term of "fake news" (which partially overlaps with disinformation and misinformation [69]) to refer to misleading information and news contents, we utilize the narrow term "disinformation" in the sense recommended by the UK government. It defines *"disinformation as the deliberate creation and sharing of false and/or manipulated information that is intended to deceive and mislead audiences, either for the purposes of causing harm, or for political, personal or financial gain"* [70]. Regarding disinformation in IJ, in a 2017 interview with Quartz, De la Peña stated that *"VR will be used for propaganda. It will be used badly for journalism. [...] But that's always going to be about, who's the maker? And it's not about the medium"* [71]. For this reason, the main concern addressed here is malevolent creativity exhibited by malicious makers. (Naturally, other concerns may stem from unintentional human errors.) As it has long been the case in cybersecurity and now also in deepfakes, there is certainly an attacker-defender arms race when it comes to disinformation attempts. Future IJ and also VR itself could arguably follow this type of trend [17].

C. Manipulated VR News and False Memory Construction

In short, with VR technologies becoming cheaper and more widespread, immersive falsehood fabricated by malicious actors could emerge in IJ settings. Sanchez described related possible dystopian scenarios *"where users are immersed in a world of fake news"* [57] while Uskali and Ikonen [72] stressed that IJ experts should be aware of *"[...] advanced and sophisticated manipulation and disinformation operations [...]"*. Beyond that, Uskali et al. specify that *"our brain believes so strongly in what it sees in VR that we might not be able to distinguish fake news from real news"* [73]. In our view, one very specific concern for the future of IJ is the targeted and tailored elicitation of false memory constructions via experiential VR news contents. On the one hand, when compared to traditional desktop displays, it is known from recent findings that immersive VR with head-mounted displays affords *more memorable* experiences by combining *"visually immersive spatial representations of data with our vestibular and proprioceptive senses"* [74]. On the other hand, this concise feature of long-lasting effects via the spatially-centered experiences in VR journalism [63] could open up a novel attack surface for malevolent actors interested in disinformation operations. More generally, Liv and Greenbaum [75] postulate that *"creating false memories to promote the uptake of fake news, both on the individual and mass scale can be enabled through multiple different means, including narrative, video, photos, and virtual reality"*. In this line, a study of Frenda et al. emphasizes that fake memories can be specifically brought forth for manipulative political gain – with successful uptakes especially if the contents are coherent with pre-existing preferences [76]. Another study found that elementary-aged children are susceptible to false memory formation in VR [77] and concluded more broadly that *"third parties may be able to elicit false memories without the consent or mental effort of an individual"*. Already the exposure to a small set of misleading photographs and a narrative led to false memory construction across half of the adult participants in a 2018 study in the period preceeding the Ireland abortion referendum [78]. Overall, it is easily conceivable that hyperrealistic IJ pieces experienced with VR headsets may exacerbate such psychological effects [79].

Against the backdrop of the foregoing analysis speaking to the creation of *durable false memories* for disinformation purposes, the following exemplary set of 3 processes could facilitate this endeavor: 1) *persuasive spatial dynamics engineering*, 2) *memory-centered sensory stimulation*, 3) *information gathering*. *Process 1* refers to any set of systematically selected processes whose outcome yields increased *spatial* awareness, perception and orientation in VR settings (such as e.g. implementing 3D minimaps [80]). *Process 2* consists in selecting any specific sensory stimulation that increases memory consolidation. For instance, it is easily conceivable that future adversaries could especially profit from the already implemented [81]–[86] but not yet available-for-sale *olfactory* displays for VR. The reason being that neuroanatomically speaking, olfactory pathways are unique [87] and olfactory memories differ from other memory forms by being particularly apt to evoke affectively-loaden memories and having a strong propensity to influence memory acquisition while being at the same time *"highly resistant to forgetting"* and *"highly resistant to retroactive interference"* [88]. (Olfactory displays can for instance be attached to VR headsets [81], [85] or utilized as on-face [86] or handheld [82] wearables.) Finally, *process 3* could include various techniques such as e.g. social engineering or open source intelligence gathering [89] (retrieving publicly available data on a target) to *identify pre-existing preferences and beliefs* of the victims to be able to match VR contents for a successful uptake of disinformation.

III. Cybersecurity-oriented Immersive Defenses

In the last Section II, we reflected upon the space of affordances available to potential malevolent actors in AI and VR respectively. We illustrated this concept utilizing the use case of disinformation in immersive journalism contexts. In this section, we discuss a cybersecurity-oriented methodology to generate defense methods against adversaries operating in AIVR, at the intersection of AI and VR. In this vein, in cybersecurity and also in recent work on security for machine learning, it is indispensable to perform a so-called *threat modelling* [31], a clear specification of assumed goals, capabilities and knowledge exhibited by the adversary. For this reason, prior to elaborating on how to generate generic immersive AIVR defense measures, we first provide a threat model for our malevolent AIVR design use case for illustration.

A. Threat Modelling for Malevolent AIVR Design Use Case

- ***Adversarial goals:*** Given the choice of our use case, the goal of the assumed adversary is a specific form of targeted disinformation by combining AI with VR tools in IJ settings. We consider that the adversary has the specific goal to manipulate the opinions, attitudes and views of selected IJ victims in a well-defined manner according to a self-defined scheme. More precisely, the goal could be to modify a source set of conceptions S to a target set of conceptions T in a certain context whereby these sets could differ in content and in confidence assigned to each element. By such a modification, the adversary intentionally aims at deceiving and misleading based on political, personal or financial motives or/and as an end in itself to cause harm. Overall, the adversarial goal would correspond to a *microtargeted disinformation* in IJ.
- ***Adversarial knowledge:*** We assume that all Generative AI components utilized are available in *white-box* settings. The same holds for the VR content creation for the IJ experiences that is fully transparent to the adversary. Moreover, the adversary is able to gain publicly available information on the victims and can attempt to gain more personal data via social engineering. One can conceive of malicious Generative AI (and by extension malicious deepfakes and VR deepfakes) as a type of adversarial examples on humans – as exposure to a specifically arranged sensorium with the goal to fool human entities (at the level of their preferences, beliefs and perceptions). Hence, in the case of humans for which information gathering succeeded in supplying crucial personal knowledge, it may be described as *grey-box* setting (a nuance between black-box and white-box adversarial knowledge levels).
- ***Adversarial capabilities:*** Regarding the Generative AI parts, the adversary can at least instrumentalize the set of 7 processes introduced in the last section which consisted of replacement, reenactment, image synthesis, speech synthesis, synthetic text generation, adversarial perturbation and automated disconcertion. In the VR content creation, the subtasks relevant to the disinformation goal are under the control of the adversary. For instance, we assume no constraints on the design and combination of the multimodal material for content (e.g. images, videos, audio samples,...). The adversary has no constraints on performing the 3 mentioned processes for VR content creation: persuasive spatial dynamics engineering, memory-centered sensory stimulation and information gathering. Thus, in total, the adversary can leverage 10 different processes to achieve microtargeted disinformation. However, it is obvious that in practice the set of capabilities could be wider and is solely constrained by malevolent creativity which is why defenses should be understood as incremental techniques and not as conclusive solutions.

B. Immersive Design Fictions for AIVR Safety

Design fiction (which we abbreviate with DF in the following) enables *"HCI and design researchers to co-create, explore and speculate the future"* [90]. Very recently, Houde et al. [28] successfully applied co-creation DF to the specific context of near-term AI safety related to (mis)use cases of Generative AI. On this basis, we regard DF as a well suited methodology for defenses against near-future AIVR safety risks as illustrated in this paper. For clarity and to facilitate a systematic procedure, we suggest to ground future AIVR DF endeavors for defenses in threat models. Moreover, the law of requisite variety in cybernetics suggests that *"only variety can destroy variety"* [91]. Applied to our use case, this signifies that in order to identify requisite knowledge for defenses against the described threat model of an adversary operating at physical, virtual and importantly immersive levels, one may profit from an immersive perspective. In our view, this need for an immersive stance for the meaningful generation of solutions applies generally to any malicious AIVR design use case linked to immersive falsehood. Interestingly, it has been proposed to utilize VR as a powerful platform for DF given the *"higher level of immersion and sense of embodiment"* [90]. In a nutshell, *AIVR safety can profit from AIVR* (next to multiple other areas such as e.g. cybersecurity, social psychology, affective science, law or journalism) and vice versa.

In the light of our threat model, it becomes clear that DFs for such malicious AIVR design use cases need to consider a *socio-psycho-technological* threat landscape with immersive, digital and physical elements and profound cognitive-affective implications. Given the complexity, a meaningful approach requires *transdisciplinary* dynamics. Importantly, the DFs need to not only address proactive defenses, but also *reactive* mechanisms [8]. In fact, proactive defenses could aim at hindering malevolent actors in AIVR to be able to disseminate their VR contents in the first place. Such measures could for instance include prevention mechanisms *preceding* content deployment and could be developed based on tools analogous to deepfake detection AI. However, given the fallibility of human knowledge, the unreliability of AI detection systems and the unpredictability of human malicious creativity, one needs to be aware of the need for reactive defense measures i.e. in the example of our use case *after* users were exposed to the manipulated VR news contents.

Notably, we do *not* consider DF as a tool to *predict* the future. Given the unpredictability of future knowledge creation, future extrapolations are limited by the state of available present knowledge and reactive measures to unknown unknowns will be needed. DF cannot foresee the consequences of not-yet created knowledge. However, DF allows the generation of plausible counterfactual paths that *could* become crucial. Organizationally, we assume a preparation phase *preceding* the DF in which an *immersive prototype* is crafted (more details below). A simple prototype could e.g. be an immersive multimodal storytelling narrative with audiovisual (see e.g. recent MIT deepfake storytelling project [92]), olfactory or tactile material. For the future, we ideally recommend a VR prototype [93]. Overall, we consider 3 disjunct groups: the makers of the immersive prototype, a set of designers with expertise in AIVR and a multidisciplinary set of participants with knowledge in a variety of technological areas overlapping with AI and VR or not. The following order for the immersive DF is non-binding and has a merely illustrative function:

1) ***Designer co-creation session:*** A group of AI and VR designers craft a *threat model 1* and a *threat model 2*. The former refers to a use case of a malicious AIVR design that would already be technically feasible nowadays and the latter to a use case that they consider feasible in 5 years given their current knowledge.
2) ***Participant introduction to AIVR:*** The AI and VR designers provide a high-level introduction to the multidisciplinary audience. It provides an overview on the state-of-the-art of technical possibilities at the intersection of AI and VR.
3) ***Designer narrative:*** The designers present *threat model 1* to the audience.
4) ***Participant co-creation session:*** Instructed by this example, the participants generate a new *threat model 3* based on what they assume might be technically feasible in 5 years given their current knowledge.
5) ***Participant narrative:*** The participants present *threat model 3* to the designers.
6) ***Narrative comparison:*** The designers present *threat model 2* and participants compare it to *threat model 3*.
7) ***Immersive session:*** Designers and participants undergo a short experience of the immersive prototype. The prototype experientially conveys a *threat model 0* (prefabricated by the makers of the prototype). In our use case example, it could consist of a short *blind* immersive experience with 2 pieces: an IJ piece (ideally in VR) featuring an *unknown but real event* and another one featuring disinformation inspired by the threat model in Subsection III-A. Before and during exposure, users are not informed on which piece is real and which manipulative. Clarification is provided at the end.
8) ***Common defense co-creation session:*** Designers, participants and makers co-create proactive *and reactive* defenses against threat models *0* to *3*. They also discuss possible adaptive attacks (when defenses are known).

IV. Conclusion

Recent research related to the safety and security of AI and VR respectively emphasizes the need to complement classical efforts to design *ethical* and safe systems with the anticipation of intentional exploits by *unethical* and malicious actors. In this vein, we performed a proactive cybersecurity-oriented analysis of malicious design in AIVR i.e. at the intersection of AI and VR. Even though the field is in its infancy, it is essential to build more robust dynamics *from the onset on* [18] and not in hindsight. By way of illustration, we applied our analysis to the use case of immersive journalism where malevolent actors could specifically harness Generative AI and VR settings for purposes of (microtargeted) disinformation creating immersive falsehood – with socio-psycho-technological implications that may require proactive and reactive *immersive defenses*.

For the purpose of generating such defense measures, we introduced a cybersecurity-oriented approach to immersive co-creation design fictions (ideally in VR). In a nutshell, *AIVR safety can benefit from immersive AIVR co-creations*. Thereby, while such co-creations may not represent a panacea to counter malicious design, it seems recommendable to incrementally employ and update them on-demand for conceivable AIVR safety use cases. Beyond that, it can be postulated that immersive design fictions inspired by security practices represent a possible way to utilize VR as rich counterfactual experiential testbed [94], [95] – however now extended to counterfactuals comprising co-existing *unethical* actors.

In a recent futures exercise, AI-generated fake content was ranked among the highest-rated potential applications for AI-enabled crime [35]. Moreover, Generative AI such as deepfakes could be used for the malicious creation of false memories [75]. Such considerations paired with the aptness of VR to facilitate durable memories represent AIVR synergies that could be exploited by malicious actors. The possible psychological implications of false memories induced in the context of such exploits could be studied in future work. Thereby, a promising avenue for future prevention and remedies could perhaps also include immersive cognitive-affective debiasing measures harnessing AIVR itself.

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.

[8] N.-M. Aliman, P. Elands, W. Hürst, L. Kester, K. R. Thórisson, 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.

[17] J. Happa, M. Glencross, and A. Steed, "Cyber security threats and challenges in collaborative mixed-reality," *Frontiers in ICT*, vol. 6, p. 5, 2019.

[18] UW Allen School Security and Privacy Research Lab, "2019 Industry-Academia Summit on Mixed Reality Security, Privacy, and Safety: Summit Report ," https://ar-sec.cs.washington.edu/research.html, 2019, online; accessed 04-August-2020.

[19] R. V. Yampolskiy, "Taxonomy of pathways to dangerous artificial intelligence," in *Workshops at the thirtieth AAAI conference on artificial intelligence*, 2016.

[20] M. Wang, X.-Q. Lyu, Y.-J. Li, and F.-L. Zhang, "VR content creation and exploration with deep learning: A survey," *Computational Visual Media*, pp. 1–26, 2020.

[21] A. J. Bose and P. Aarabi, "Virtual Fakes: DeepFakes for Virtual Reality," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2019, pp. 1–1.

[22] N. De la Peña, P. Weil, J. Llobera, E. Giannopoulos, A. Pomés, B. Spanlang, D. Friedman, M. V. Sanchez-Vives, and M. Slater, "Immersive journalism: immersive virtual reality for the first-person experience of news," *Presence: Teleoperators and virtual environments*, vol. 19, no. 4, pp. 291–301, 2010.

[23] J. V. Pavlik, *Journalism in the age of virtual reality: How experiential media are transforming news*. Columbia University Press, 2019.

[24] M. Blythe, J. Steane, J. Roe, and C. Oliver, "Solutionism, the game: design fictions for positive aging," in *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 2015, pp. 3849–3858.

[25] M. Blythe, E. Encinas, J. Kaye, M. L. Avery, R. McCabe, and K. Andersen, "Imaginary design workbooks: Constructive criticism and practical provocation," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–12.

[26] E. Cheon and N. M. Su, "Futuristic autobiographies: Weaving participant narratives to elicit values around robots," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 388–397.

[27] A. Rapp, "Design fictions for learning: A method for supporting students in reflecting on technology in Human-Computer Interaction courses," *Computers & Education*, vol. 145, p. 103725, 2020.

[28] S. Houde, V. Liao, J. Martino, M. Muller, D. Piorkowski, J. Richards, J. Weisz, and Y. Zhang, "Business (mis) Use Cases of Generative AI," *arXiv preprint arXiv:2003.07679*, 2020.

[29] A. G. Pillai, N. Ahmadpour, S. Yoo, A. B. Kocaballi, S. Pedell, V. P. Sermuga Pandian, and S. Suleri, "Communicate, Critique and Co-create (CCC) Future Technologies through Design Fictions in VR Environment," in *Companion Publication of the 2020 ACM on Designing Interactive Systems Conference*, 2020, pp. 413–416.

[30] D. H. Cropley, J. C. Kaufman, and A. J. Cropley, "Malevolent creativity: A functional model of creativity in terrorism and crime," *Creativity Research Journal*, vol. 20, no. 2, pp. 105–115, 2008.

[31] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin, "On evaluating adversarial robustness," *arXiv preprint arXiv:1902.06705*, 2019.

[32] R. S. S. Kumar, D. O. Brien, K. Albert, S. Viljöen, and J. Snover, "Failure modes in machine learning systems," *arXiv preprint arXiv:1911.11034*, 2019.

[33] C. T. Thanh and I. Zelinka, "A Survey on Artificial Intelligence in Malware as Next-Generation Threats," in *Mendel*, vol. 25, no. 2, 2019, pp. 27–34.

[34] M. Comiter, "Attacking artificial intelligence: AI's security vulnerability and what policymakers can do about it," *Belfer Center for Science and International Affairs, Harvard Kennedy School*, 2019.

[35] M. Caldwell, J. Andrews, T. Tanay, and L. Griffin, "AI-enabled future crime," *Crime Science*, vol. 9, no. 1, pp. 1–13, 2020.

[36] L. Blackwell, N. Ellison, N. Elliott-Deflo, and R. Schwartz, "Harassment in social virtual reality: Challenges for platform governance," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–25, 2019.

[37] F. O'Brolcháin, T. Jacquemard, D. Monaghan, N. O'Connor, P. Novitzky, and B. Gordijn, "The convergence of virtual reality and social networks: threats to privacy and autonomy," *Science and engineering ethics*, vol. 22, no. 1, pp. 1–29, 2016.

[38] S. Cole and E. Maiberg, "Deepfake Porn Is Evolving to Give People Total Control Over Women's Bodies," https://www.vice.com/en_uk/article/9keen8/deepfake-porn-is-evolving-to-give-people-total-control-over-womens-bodies, 2019, online; accessed 04-August-2020.

[39] T. Macaulay, "New AR app will let you model a virtual companion on anyone you want," https://thenextweb.com/neural/2020/06/01/new-ar-app-will-let-you-model-a-virtual-companion-on-anyone-you-want/, 2020, online; accessed 04-August-2020.

[40] S. Srivastava, "Analysing the Futuristic Potentials of Deepfake + Augmented and Virtual Reality," https://www.analyticsinsight.net/analysing-the-futuristic-potentials-of-deepfake-augmented-and-virtual-reality/, 2020, analytics Insight; accessed 04-August-2020.

[41] N. Caporusso, "Deepfakes for the Good: A Beneficial Application of Contentious Artificial Intelligence Technology," in *International Conference on Applied Human Factors and Ergonomics*. Springer, 2020, pp. 235–241.

[42] M. Albahar and J. Almalki, "Deepfakes: Threats and countermeasures systematic review," *Journal of Theoretical and Applied Information Technology*, vol. 97, no. 22, pp. 3242–3250, 2019.

[43] Y. Mirsky and W. Lee, "The Creation and Detection of Deepfakes: A Survey," *arXiv preprint arXiv:2004.11138*, 2020.

[44] A. E. Venema and Z. J. Geradts, "Digital Forensics, Deepfakes, and the Legal Process," *TheSciTechLawyer*, vol. 16, no. 4, pp. 14–23, 2020.

[45] N. Carlini and H. Farid, "Evading deepfake-image detectors with white- and black-box attacks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 658–659.

[46] R. Satter, "Deepfake Used to Attack Activist Couple Shows New Disinformation Frontier," https://www.reuters.com/article/us-cyber-deepfake-activist/deepfake-used-to-attack-activist-couple-shows-new-disinformation-frontier-idUSKCN24G15E, 2020, Reuters; accessed 04-August-2020.

[47] ——, "Experts: Spy used AI-generated face to connect with targets," https://apnews.com/bc2f19097a4c4fffaa00de6770b8a60d, 2019, Associated Press (AP); accessed 04-August-2020.

[48] Y. Mirsky, T. Mahler, I. Shelef, and Y. Elovici, "CT-GAN: Malicious tampering of 3D medical imagery using deep learning," in *28th USENIX Security Symposium (USENIX Security 19)*, 2019, pp. 461–478.

[49] C. Stupp, "Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case," https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402, 2019, The Wall Street Journal; accessed 04-August-2020.

[50] P. Tully and L. Foster, "Repurposing Neural Networks to Generate Synthetic Media for Information Operations," https://www.blackhat.com/us-20/briefings/schedule/#repurposing-neural-networks-to-generate-

synthetic-media-for-information-operations-19529, 2020, Session at blackhat USA 2020; accessed 08-August-2020.
[51] P. Neekhara, S. Hussain, M. Jere, F. Koushanfar, and J. McAuley, "Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples," *arXiv preprint arXiv:2002.12749*, 2020.
[52] Q. Sun, A. Tewari, W. Xu, M. Fritz, C. Theobalt, and B. Schiele, "A hybrid model for identity obfuscation by face replacement," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 553–569.
[53] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging Frequency Analysis for Deep Fake Image Recognition," *arXiv preprint arXiv:2003.08685*, 2020.
[54] K. Hao, "The Biggest Threat of Deepfakes Isn't the Deepfakes Themselves," *MIT Technology Review*, 2019.
[55] A. Riikonen, "Decide, Disrupt, Destroy: Information Systems in Great Power Competition with China." *Strategic Studies Quarterly*, vol. 13, no. 4, 2019.
[56] L. Knoepp, "Forget Oculus Rift, Meet The Godmother Of VR," https://www.forbes.com/sites/lillyknoepp/2017/04/13/forget-oculus-rift-meet-the-godmother-of-vr/, 2017, Forbes; accessed 04-August-2020.
[57] A. L. Sánchez Laws, "Can immersive journalism enhance empathy?" *Digital Journalism*, vol. 8, no. 2, pp. 213–228, 2020.
[58] J. V. Pavlik, "Drones, augmented reality and virtual reality journalism: Mapping their role in immersive news content," *Media and Communication*, vol. 8, no. 3, pp. 137–146, 2020.
[59] M. Bujić, M. Salminen, J. Macey, and J. Hamari, ""Empathy machine": how virtual reality affects human rights attitudes," *Internet Research*, 2020.
[60] A. Dhanda, M. Reina Ortiz, A. Weigert, A. Paladini, A. Min, M. Gyi, S. Su, S. Fai, and M. Santana Quintero, "RECREATING CULTURAL HERITAGE ENVIRONMENTS FOR VR USING PHOTOGRAMMETRY," *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLII-2/W9, pp. 305–310, 2019.
[61] M. Obradović, I. Vasiljević, I. DJurić, J. Kićanović, V. Stojaković, and R. Obradović, "Virtual Reality Models Based on Photogrammetric Surveys – A Case Study of the Iconostasis of the Serbian Orthodox Cathedral Church of Saint Nicholas in Sremski Karlovci (Serbia)," *Applied Sciences*, vol. 10, no. 8, p. 2743, 2020.
[62] S. Kang, E. O'Brien, A. Villarreal, W. Lee, and C. Mahood, "Immersive Journalism and Telepresence: Does virtual reality news use affect news credibility?" *Digital Journalism*, vol. 7, no. 2, pp. 294–313, 2019.
[63] M. Cadoux, "AR and VR will make spatial journalism the future of reporting," https://venturebeat.com/2019/11/10/ar-and-vr-will-make-spatial-journalism-the-future-of-reporting/, year = 2019, note = VentureBeat; accessed 04-August-2020.
[64] G. M. Hardee and R. P. McMahan, "FIJI: a framework for the immersion-journalism intersection," *Frontiers in ICT*, vol. 4, p. 21, 2017.
[65] M. Bujić and J. Hamari, "Immersive journalism: Extant corpus and future agenda," *CEUR-WS*, 2020.
[66] R. Mabrook, "Collaborative and Experimental Cultures in Virtual Reality Journalism: From the Perspective of Content Creators," *International Journal of Humanities and Social Sciences*, vol. 13, no. 5, pp. 532–542, 2019.
[67] I. Tribusean, "The Use of VR in Journalism: Current Research and Future Opportunities," in *Augmented Reality and Virtual Reality*. Springer, 2020, pp. 227–239.
[68] D. G. Johnson, "Promises and perils in immersive journalism," *Immersive Journalism as Storytelling: Ethics, Production, and Design*, 2020.
[69] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild *et al.*, "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.
[70] D. Colins, "Disinformation and "Fake News": Interim Report: Government Response to the Committee's Fifth Report of Session 2017–19," *UK House of Commons Digital*, 2018.
[71] QUARTZ, "Virtual reality, fake news and the future of fact," https://www.youtube.com/watch?v=i5LW03vw_x8, 2017, YouTube video; accessed 04-August-2020.
[72] T. Uskali and P. Ikonen, "THE IMPACT OF EMOTIONS IN IMMERSIVE JOURNALISM," *Immersive Journalism as Storytelling: Ethics, Production, and Design*, 2020.
[73] T. Uskali, A. Gynnild, S. Jones, and E. Sirkkunen, *Immersive Journalism as Storytelling: Ethics, Production, and Design*. Routledge, 2020.
[74] E. Krokos, C. Plaisant, and A. Varshney, "Virtual memory palaces: immersion aids recall," *Virtual Reality*, vol. 23, no. 1, pp. 1–15, 2019.
[75] N. Liv and D. Greenbaum, "Deep Fakes and Memory Malleability: False Memories in the Service of Fake News," *AJOB neuroscience*, vol. 11, no. 2, pp. 96–104, 2020.
[76] S. J. Frenda, E. D. Knowles, W. Saletan, and E. F. Loftus, "False memories of fabricated political events," *Journal of Experimental Social Psychology*, vol. 49, no. 2, pp. 280–286, 2013.
[77] K. Y. Segovia and J. N. Bailenson, "Virtually true: Children's acquisition of false memories in virtual reality," *Media Psychology*, vol. 12, no. 4, pp. 371–393, 2009.
[78] G. Murphy, E. F. Loftus, R. H. Grady, L. J. Levine, and C. M. Greene, "False memories for fake news during Ireland's abortion referendum," *Psychological science*, vol. 30, no. 10, pp. 1449–1459, 2019.
[79] M. Slater, C. Gonzalez-Liencres, P. Haggard, C. Vinkers, R. Gregory-Clarke, S. Jelley, Z. Watson, G. Breen, R. Schwarz, W. Steptoe *et al.*, "The ethics of realism in virtual and augmented reality," *Frontiers in Virtual Reality*, vol. 1, p. 1, 2020.
[80] J. Kotlarek, I.-C. Lin, and K.-L. Ma, "Improving spatial orientation in immersive environments," in *Proceedings of the Symposium on Spatial User Interaction*, 2018, pp. 79–88.
[81] T. Nakamoto, T. Hirasawa, and Y. Hanyu, "Virtual environment with smell using wearable olfactory display and computational fluid dynamics simulation," in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2020, pp. 713–720.
[82] S. Niedenthal, P. Lundén, M. Ehrndal, and J. K. Olofsson, "A handheld olfactory display for smell-enabled VR games," in *2019 IEEE International Symposium on Olfaction and Electronic Nose (ISOEN)*. IEEE, 2019, pp. 1–4.
[83] J. M. M. Martins and M. de Paiva Guimarães, "Using Olfactory Stimuli in Virtual Reality Applications," in *2018 20th Symposium on Virtual and Augmented Reality (SVR)*. IEEE, 2018, pp. 57–64.
[84] G. Tsaramirsis, M. Papoutsidakis, M. Derbali, F. Q. Khan, and F. Michailidis, "Towards Smart Gaming Olfactory Displays," *Sensors*, vol. 20, no. 4, p. 1002, 2020.
[85] J. A. Raines and D. E. Litt, "Olfactory simulation system for head-mounted displays," Apr. 30 2020, uS Patent App. 16/670,572.
[86] Y. Wang, J. Amores, and P. Maes, "On-Face Olfactory Interfaces," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–9.
[87] R. M. Sullivan, D. A. Wilson, N. Ravel, and A.-M. Mouly, "Olfactory memory networks: from emotional learning to social behaviors," *Frontiers in behavioral neuroscience*, vol. 9, p. 36, 2015.
[88] H. L. Roediger, F. M. Zaromb, and W. Lin, "1.02 - A Typology of Memory Terms," in *Learning and Memory: A Comprehensive Reference (Second Edition)*, 2nd ed., J. H. Byrne, Ed. Oxford: Academic Press, 2017, pp. 7 – 19. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780128093245210031
[89] L. D. Ball, G. Ewan, and N. J. Coull, "Undermining: social engineering using open source intelligence gathering," in *4th International Conference on Knowledge Discovery and Information Retrieval*. Scitepress Digital Library, 2012, pp. 275–280.
[90] N. Ahmadpour, S. Pedell, A. Mayasari, and J. Beh, "Co-creating and assessing future wellbeing technology using design fiction," *She Ji: The Journal of Design, Economics, and Innovation*, vol. 5, no. 3, pp. 209–230, 2019.
[91] W. R. Ashby, *An introduction to cybernetics*. Chapman & Hall Ltd, 1961.
[92] MIT Open Learning, "Tackling the misinformation epidemic with "In Event of Moon Disaster" ," https://news.mit.edu/2020/mit-tackles-misinformation-in-event-of-moon-disaster-0720, 2020, MIT News; accessed 11-October-2020.
[93] M. Stepanovic and V. Ferraro, "Reflecting on New Approaches for the Design for Behavioural Change Research and Practice: Shaping the Technologies Through Immersive Design Fiction Prototyping," in *International Conference on Human-Computer Interaction*. Springer, 2020, pp. 542–560.
[94] N.-M. Aliman and L. Kester, "Extending Socio-Technological Reality for Ethics in Artificial Intelligent Systems," in *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE, 2019, pp. 275–2757.
[95] ——, "Transformative AI governance and AI-Empowered ethical enhancement through preemptive simulations," *Delphi Interdisc. Rev. Emerg. Technol*, vol. 2, no. 1, pp. 23–29, 2019.