Exploring User Reactions and Mental Models Towards Perceptual Manipulation Attacks in Mixed Reality

Kaiming Cheng, Jeffery F. Tian, Tadayoshi Kohno, and Franziska Roesner, *University of Washington*

https://www.usenix.org/conference/usenixsecurity23/presentation/cheng-kaiming

This paper is included in the Proceedings of the 32nd USENIX Security Symposium.

August 9–11, 2023 • Anaheim, CA, USA

978-1-939133-37-3

Open access to the Proceedings of the 32nd USENIX Security Symposium is sponsored by USENIX.

Exploring User Reactions and Mental Models Towards Perceptual Manipulation Attacks in Mixed Reality

Kaiming Cheng, Jeffery F. Tian, Tadayoshi Kohno, Franziska Roesner
Paul G. Allen School of Computer Science & Engineering, University of Washington
https://ar-sec.cs.washington.edu
{kaimingc, jefftian, yoshi, franzi}@cs.washington.edu

Abstract

Perceptual Manipulation Attacks (PMA) involve manipulating users' multi-sensory (e.g., visual, auditory, haptic) perceptions of the world through Mixed Reality (MR) content, in order to influence users' judgments and following actions. For example, a MR driving application that is expected to show safety-critical output might also (maliciously or unintentionally) overlay the wrong signal on a traffic sign, misleading the user into slamming on the brake. While current MR technology is sufficient to create such attacks, little research has been done to understand how users perceive, react to, and defend against such potential manipulations. To provide a foundation for understanding and addressing PMA in MR, we conducted an in-person study with 21 participants. We developed three PMA in which we focused on attacking three different perceptions: visual, auditory, and situational awareness. Our study first investigates how user reactions are affected by evaluating their performance on "microbenchmark" tasks under benchmark and different attack conditions. We observe both primary and secondary impacts from attacks, later impacting participants' performance even under non-attack conditions. We follow up with interviews, surfacing a range of user reactions and interpretations of PMA. Through qualitative data analysis of our observations and interviews, we identify various defensive strategies participants developed, and we observe how these strategies sometimes backfire. We derive recommendations for future investigation and defensive directions based on our findings.

1 Introduction

Mixed reality (MR) technologies — technologies that place virtual content in users' real-world environment — are poised to dramatically alter how people interact with the physical world, the digital world, and each other.[1] Once inaccessible to the general public, MR devices are becoming more available and affordable. Technologies like Microsoft's Hololens 2 [3], Meta's Oculus Quest 2 [6], Project Cambria [5], Snapchat's Spectacles [11], and Apple's incoming MR headset [1] are transforming previous visions for MR into reality.

Despite the benefit MR technologies could deliver, a growing body of research in the computer security and MR communities has looked at perceptual issues in MR [21, 36] and how users could be manipulated by content created by MR applications. One class of potential attacks, termed *Perceptual Manipulation Attacks* (PMA) by Tseng et al. [68], aims to manipulate the human multi-sensory perceptions of the physical world to influence users' decision-making and even lead to physical harm through the presented MR output stimuli. Unlike attacks targeting vulnerable hardware or platforms, here the attack is impacting the perception and/or cognition of a person in an immersive way using the MR system. While PMA also exist on traditional platforms (e.g., phishing, distracting pop-ups), the MR experience is fundamentally more immersive: for example, previous studies in gaming settings [16, 48, 72] suggested that MR might produce meaningfully different experiences, such as higher presence, more real and personal involvement, and higher affective responses.

PMA in MR are not only a theoretical threat; precursors are starting to manifest in practice. Although not adversarial in nature, recent research [35, 70] documented severe negative psychological impacts and physical injuries including even death on users from using real-world MR applications (such as Pokémon Go), which indicates that even benignly-designed MR applications can unintentionally "hack" user perceptions and introduce critical risks. Prior work also demonstrates different types of PMA in a laboratory setting, such as obscuring important real-world content [40], creating audio indistinguishable from reality [4], alternating perceived haptic softness level [49, 50, 66], affecting users' gustatory sensations [45], and disrupting users physiologically [15].

Thus, while it is clear that PMA are possible, both with today's technologies and in the future, what is not known are the human experiences while undergoing such attacks, where "experiences" include both behaviors (e.g., actions) and thoughts

[1] We use the term "mixed reality (MR)" to refer to technologies that place virtual content in a real-world environment, whether embedded in it or overlaid on it. Other works may use other terms to refer to the same or related concepts, including augmented reality (AR), extended reality (XR), and (pass-through) Virtual Reality (VR).

(e.g., impressions, interpretations). As our community seeks to develop defenses against PMA in MR environments, we argue that it is essential to understand users' experiences with PMA. Such an understanding can form the basis of future risk assessments and defensive approaches.

Research Questions. Motivated by the above, we formulate the following two key research questions:

1. *RQ1:* What physical or behavioral reactions and responses do users have when experiencing perceptual manipulation attacks (PMA) in MR?
2. *RQ2:* What are user-reported reflections, reactions, and defensive strategies to PMA in MR during or shortly after they occur? For example: Do users rationalize their behaviors and responses to attacks? Do they perceive that they are under attack during the attack? How do users defend against such attacks?

Foreshadowing to Section 7, answering the above questions enable us to suggest key strategies for researchers and industry to reduce the harms of PMA in MR in the wild. These strategies include both preventative measures and approaches for resiliency and harm reduction if attacks manifest.

Methodology. To answer RQ1 and RQ2, we develop a methodology with the following approach: we subject participants to PMA and (1) observe their physical and behavioral reactions and responses during the attacks and (2) listen to any thoughts they might express during the attacks and interview them after.

We highlight here several key elements of our methodology. First, because we would be subjecting participants to PMA in a laboratory setting, it was essential to design a safe testing procedure, including but not limited to receiving IRB approval. Second, it was essential to create an experimental apparatus that (A) exposed participants to programmatic MR content (virtual content placed in the physical world), (B) allowed that content to sometimes (but not always) be adversarial, and (C) control the physical environment such that the experiments would be repeatable. Elaborating on (C), it was essential for the physical environment and the MR content (including adversarial content) to be synchronized.

To meet our experimental objectives, we designed an experimental harness with the following properties: The physical world environment included a computer monitor and mouse. The participants wore a mixed-reality headset. When the participants looked at the real-world computer monitor while wearing the headset, they saw that monitor. Our experimental harness advanced both the content displayed on the computer monitor and the virtual content displayed within the mixed-reality headset. See Figure 1.

We experimented with three controlled scenarios in which we asked participants to do tasks where their performance was measured using three cognitive metrics: reaction speed, sustained attention, and focus. Our study was a deception study: participants did not initially know that we would be subjecting them to PMA. Within these scenarios, we then subjected participants to three different PMA, each with the intention to attack different perceptions: visual, auditory, and situational awareness.

Figure 1: Researcher testing our experimental harness.

Contributions. In summary, our contributions include:

1. *End User Behavioral Reactions:* We focus on end users' *experiences* when they encounter adversarial MR content. Our results suggest perceptual manipulation attacks (PMA) successfully disrupt user performance and evoke the adversarially intended reaction from users — as well as secondary effects, such as slowing down on non-attack settings or amplifying subsequent PMA impacts.
2. *End User Self-Reported Reflections:* Through follow-up interviews, we provide rich qualitative insights about how people assess, reason about, and defend against PMA in MR in practice.
3. *MR PMA Experimental Harness:* We implemented a harness to capture the impact of PMA in MR on end users. We have publicly released our experimental harness[2] to facilitate open science.
4. *Foundation for Future Defenses:* Stepping back and reflecting on our studies of PMA with real users, we provide suggestions for researchers and industry to build the next generation MR defenses and strategies to reduce the harms of attacks in the wild.

2 Background and Related Work

2.1 Security and Privacy Challenges for Mixed Reality

Broadening from the concept of augmented reality (AR), which dates at least to Sutherland in the 1960s [65], Milgram & Kishin [43] introduced the concept of *mixed reality* (MR) in 1994 and defined it as any environment that blends real and virtual visual objects. Today, MR can refer to extending all types of perception, including audio, motion, haptics, taste, and smell. For the past two decades, researchers in MR have focused on delivering the necessary technology and exploring various application possibilities, including education, fabrication, entertainment, surgery, and training [17,31,32,59,71,76].

[2] https://github.com/UWCSESecurityLab/MR-PMA-Harness

As MR technologies rapidly evolve, the computer security community and MR industry have begun to identify key security and privacy challenges in this space (e.g., [54]). Roesner et al. [55] first proposed security threat modeling taxonomies for AR, which included input, data access, and output; Guzman et al. [20] extended these three aspects to include user interaction and device protection. Much prior MR security work falls into these taxonomies, including work focusing on the privacy of sensor data input to MR platforms [24, 27, 33, 34, 69, 77], and work addressing output security: preventing apps from displaying unwanted or malicious content [14, 40]. Other works consider more traditional notions of security for MR, such as authentication and network security [29, 63], and others explore secure multi-user collaborative interactions [41, 53, 57]. Prior works also proposed mitigating strategies on the MR system level [34, 39, 40, 56, 57, 75].

2.2 Impact on Perception from External Stimuli

Prior work in non-MR contexts already showed how human cognition and perception can be distracted or manipulated by exogenous cues (i.e., external stimuli). For example, a line of research [23, 30, 38, 67] studied how visual reaction time is sensitive to visual stimuli. Yantis and Jonides [74] showed that an object with sudden onset was always processed first, which is related to our study design in Section 4.1. Neyens and Boyle [46] suggested that cell phone usage while driving is associated with cognitive, auditory, and visual distractions, causing a high likelihood of vehicle accidents, which is related to our study design in Section 4.2. Simons and Chabris conducted the famous *attentional blindness experiment* in 1999 [62]. When asked to perform a task that required full attentional resources, subjects often failed to see a gorilla in the midst of the experiment, which is related to our study design in Sec tion4.3. While the above work has studied the impact of external stimuli on perception, and provides theoretical support for our qualitative results, our work dives deeper into the experiences of users encountering these techniques deployed by an adversary in an immersive, MR space. We suggest mitigation strategies based on our rich qualitative insights.

2.3 Perceptual Manipulations in Mixed Reality

MR, given its immersive nature, can be an even more powerful medium for perceptual manipulation. Previous work has explored different techniques in MR to manipulate various kinds of human perception. Schmidt et al. [60] leverage visual illusions to manipulate the perceived spatial relationships between the user and objects in MR. Nakano et al. [45] developed a generative adversarial network based MR application that changes the appearance of food in order to manipulate users' gustatory sensations. Punpongsanon et al. [49, 50] investigated how MR visual output can affect human perception of haptic softness and bending stiffness. Researchers have also developed techniques that manipulate users' visual perception to imperceptibly redirect their movement in the physical space [37, 64].

Recently, security researchers started to explore the potential of attacks based on perceptual manipulation. Baldassi et al. [15] considered direct impacts on the human brain, identifying sensory and perceptual risks (e.g., from accidentally or maliciously induced visual adaptations, motion-induced blindness, and photosensitive epilepsy). Casey et al. [19] present several proof-of-concept attacks that manipulate user visual perception to direct their physical movement, collide with real world objects, and induce motion sickness. Tseng et al. [68] identify a set of harmful scenarios using PMA through speculative design workshops.

Though a growing body of work has now raised awareness of perceptual manipulation in MR and associated security issues, almost none of the work has *empirically studied* user mental models when experiencing such attacks. The lack of prior studies is explainable because MR systems are not yet widely deployed. However, we consider a study of users' reactions to and perceptions of of PMA in advance of their manifestation in practice to be important. We aim to begin closing this gap in our work here.

3 Methodology

To empirically explore user reactions to PMA, we designed an in-lab study in a user study room. As discussed in Section 1, ours was a deception study in which participants were told that we were evaluating the impact of wearing a MR headset while conducting tasks. We mounted PMA on participants as they performed these tasks, interviewed them, and then debriefed them after completing all tasks. Each study lasted for around 60 minutes.

3.1 Study Procedure

Warm-up Phase. We started by having participants familiarize themselves with the MR world. We helped participants, by non-contact demonstration, adjust the headset and tune the interpupillary distance to minimize their discomfort level.

Experiment Phase. Participants completed specified tasks in a benchmark setting and under the influence of PMA. We intentionally did not mention security to them until a debriefing at the end of the study, to minimize priming participants to the possibility of attacks. Then, participants were instructed to complete three tasks (see Section 4 for details). Most tasks consisted of three rounds: (1) an initial *task-training* round without the headset, followed by (2) a *benchmark* round where they completed the task while wearing the headset (though no MR content was shown), followed by (3) a *withAttack* round during which we mounted PMA while they completed the task. We measured and recorded participants' performance on the tasks under each condition. After the benchmark round, we did not give instructions about any virtual content in order to observe the participants' organic mental models; the task goal remained the same after the benchmark round.

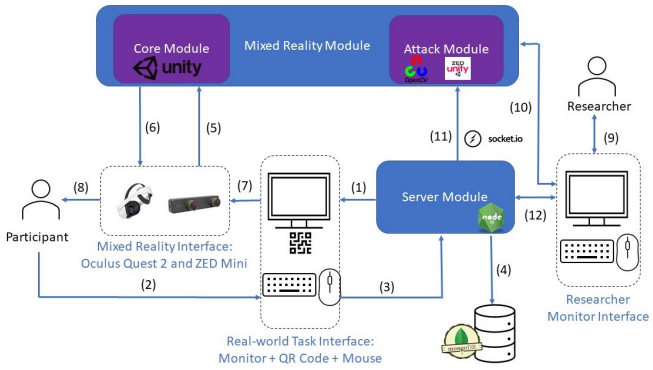


Figure 2: Diagram of our experimental harness.

We encouraged participants to talk aloud throughout the experiment to capture in-the-moment reactions. For all experiments, researchers could see participants' view through the headset streamed to a separate desktop computer display. With participant consent, we recorded the entire study, including this captured first-person view; we present screenshots of this view in figures throughout the paper.

We emphasize that our *benchmark* round includes *no* MR content at all, rather than attempting to compare with a condition containing benign MR output. There are many ways that a benign or intending-to-be helpful MR application could be designed, and a benign application might accidentally influence user perception as well (as in Pokémon Go). Thus, in our work, we focus on exploring user reactions to an *intentionally* manipulative MR application, compared to no MR content at all; future experiments could explore the spectrum of reactions user might have to non-adversarial MR content as well.

Post-Task Interview Phase. We conducted an in-depth, qualitative interview with participants. We asked questions about their experiences doing the tasks, beginning with more open-ended questions to avoid priming them. We then asked follow-up questions if participants discussed the adversarial MR outputs, and we eventually debriefed participants and disclosed the purpose of the study. We continued to talk to participants for about 15 minutes about their reflections on the study in particular and PMA in general (see Appendix B). We do not include any information from these post-debrief conversations in the results because participants had been primed about PMA at this point, but we believe that this post-study session helped participants process and understand the study. Before participants left the room, we reminded them that they could reach out to us if they felt discomfort or experienced any negative impact from the study. No one mentioned, nor did we observe, any concern or discomfort. We include our interview script as well as our debrief script in Appendix B.

3.2 Experimental Harness

Figure 2 presents an overview of our experimental harness. The computer and mouse correspond to the real-world tasks presented to users. Affixed to the monitor is a QR code, which enables the localization of the PMA output. In the following, numbers refer to the arrows in Figure 2.

The Server Module controls the benchmark experiment. It (1) determines what content to display on the monitor (content corresponding to the real-world task that the user is performing). It also receives user input (2), in the form of mouse clicks (3), and saves user's performance in our MongoDB database (4). The Server Module is implemented with Nodejs. The Server Module, combined with the monitor and the mouse, is the entirety of the participant interface during the *task-training* round of the experimental phase.

The Mixed Reality Interface uses the Oculus Quest 2 head-mounted display with a ZED Mini camera. We attached the ZED Mini to the Oculus headset with adhesive tape. The Unity-based Mixed Reality Module renders the ZED Mini camera stereo view (5) in the Oculus (6), and outputs it with the experiment (7) to user (8) and researcher (9). This part of the Mixed Reality Module, along with the Server Module, constitute the entirety of the participant interface during the *benchmark* round of the experimental phase.

During the *withAttack* round of each scenario, the researchers start the Mixed Reality Module, as they did during the *benchmark* round (9 and 10). The Server Module sends the trigger via Socket.io to the Attack Module (11). The Attack Module leverages the OpenCV library to detect the QR code (via (5)) on the Task Interface. It calculates the adversarial output's placement and generates it. The MR module then mixes the adversarial MR output with the ZED Mini input (5) to render visual and auditory output (6) and displays it to user (8), and to researchers (10 and 9).

3.3 Recruitment and Screening

Due to COVID-19, university safety protocols, as well as the nature of MR requiring participants to be in person, we recruited participants with access to school buildings via department Slack and personal contacts. A study session took around 80 minutes including sanitizing the equipment.

We advertised the study goal as evaluating the impact of wearing a MR headset while conducting a primary task: we did not advertise it as a MR attack study to minimize priming participants and potentially affecting their behaviors.

Candidates completed a screening survey (Appendix A), indicating any previous AR/MR/VR experience, demographics, and contact information (name and email). We did not consider for the full study anyone who indicated dizziness or nausea during previous AR/MR/VR use. Ultimately, we recruited 21 participants (10 men, 11 women; age: $M = 22.12$, $SD = 4.31$). Overall, most participants (around 85%) had "some" experience with AR/MR/VR, two participants are regular users, and one participant had never tried AR/MR/VR.

Our participants' ages ranged from 20 to 28; the majority are in the technology industry.

Participants who completed the full study were compensated with a $30 Amazon gift card. The compensation was based approximately on the hourly minimum wage in our area ($15) and accounting for additional time that participants might need to commute to our lab. The compensation method and amount were approved as part of our full IRB protocol. We note that compensation may influence a participant's decision to participate in the study in the first place.

3.4 Ethical Considerations

Our study, reviewed and approved by our university's IRB, raised several ethical considerations. First, it was designed as a deception study: we did not initially disclose its true goals so we could avoid biasing reactions. We designed the study to minimize any potential risks beyond task performance impairments: participants completed the tasks while seated, and the tasks did not involve moving around the physical space while wearing the MR headset. We informed participants that the MR headset could cause discomfort (such as nausea or dizziness) and that they could stop at any time during the study with no loss of promised compensation. (No participants discontinued the study before completion, and none expressed or appeared to experience discomfort during the study.) Since attacks could affect participant performance on tasks, we framed the study as an evaluation of the MR technology rather than participant performance. We debriefed all participants about the true nature of the study at the end.

Additionally, we did not ask participants to reveal sensitive information. We sent the consent form to participants days before the study and asked for their physical signature when they arrived at the user study room to participate in the study and to be video recorded. We stored all recordings on password-protected drives and removed any personally identifying information from notes and transcripts.

3.5 Data Analysis

We analyzed our data using both quantitative and qualitative methods. In the experiment phase, we evaluated participants' performance for each task under various conditions using the metrics we describe in Section 4. For the qualitative data, we transcribed the interview audio using Rev [8]. All four researchers independently developed preliminary codebooks based on three interviews. These researchers then iteratively resolved disagreements and developed a full version of the codebook collaboratively. The first author applied the codebook to all interview transcripts. All researchers discussed when new codes emerged, and resolved any further disagreements. The first author kept the codebook updated, and applied the final codebook to all interview transcripts. No new theme was found. We conducted thematic analyses from a broad family of methods [18, 42], combining a deductive approach (applying a security threat modeling framework to our interview data to identify sub-themes related to attack attribution and defensive strategy) and an inductive approach (generating additional themes/codes from the interview data). We provide the full codebook in Appendix C.

3.6 Limitations

First, though we report experimental data for participants' performance, we do not aim to make causal or generalizable claims. We did not conduct a large-scale randomized trial with many participants and, for example, our experiments do not account for possible ordering effects. While we asked participants to randomly select either Scenario A or Scenario B to start (see Section 4), Scenario C always came last because it most clearly gave away the adversarial nature of the study. Our work is the first step towards deeply understanding user perceptions and defensive approaches to PMAs, and lays a foundation for future work to conduct larger-scale studies to test ordering effects or the generality of our findings.

Second, our participant sample has the following limitations: Most of our participants were predominantly young adults with STEM education backgrounds. Due to these limitations, the findings of our study should not be generalized. Future work could explore a broader population or more directly focus on specific populations.

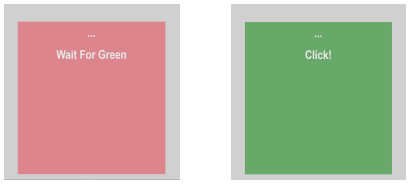
Third, because participants were in a lab setting and in some cases previously knew the researchers, they may have either trusted the MR content more (assuming good intentions of the researchers) or less (if participants happened to know that the researchers work in a security-focused group). To mitigate these potential impacts, we designed the study to avoid priming participants about security. We found that participants *were* impacted by the adversarial MR content in our experiments, even in cases where they assumed good intentions (e.g., attributing attacks to glitches) or knew the researchers. Likewise, we expect that users in real MR scenarios will bring a variety of preconceptions to the situation.

Fourth, we did not ask participants about color blindness during our screening (though we should have). No participants mentioned color-blindness during the tasks, and their task performance suggests that they could see the colors we used.

Finally, we conducted our experiment with one particular hardware setup. We chose state-of-the-art hardware and software, but our results may not generalize to other setups or future technologies. Despite the imperfections of the MR setup (e.g., virtual content did not necessarily believably blend into the real world), participants were impacted by our attacks.

4 Scenarios and Attacks

Our high-level goal is to explore how users are impacted by and respond to different PMA in MR. For this investigation, we thus design and develop controllable, repeatable in-lab experiments that were modeled after both known psychological experiments and prior concern about MR in real environments. In choosing our attacks, we focus on exploring



(a) Waiting Page (b) Reaction Page

Figure 3: Participant view of the real-world Reaction Task.

different types of perceptions to attack: visual, auditory, and situational awareness. Note that the attacks we present are prototypes; future attackers might mount far more sophisticated and powerful attacks with the help of next generation MR headsets and external devices such as eye tracking hardware [12] or electromyography wearable wristbands [2].

The following three subsections describe our three scenario case studies. We summarize all of our attacks in Table 1.

4.1 Scenario A: Reaction Time Task, Visual Attacks

Reaction Time Task. Reaction time is the duration of the interval between presentation of a stimulus and response to the stimulus. There are concerns about how adversarial MR content might manipulate user visual perception to impact those reaction times rooted in classical psychology literature [28,47]. For example, for people using MR while driving, attackers could overlay virtual objects on real traffic lights, causing the driver to react slowly or to misinterpret those lights [25,26]. To evaluate the effect of adversarial MR output on reaction time, we created an experiment in which participants were asked to respond quickly to a stimulus, modeled after an existing cognitive study game [10]. The real-world task consisted of red and green boxes shown on a computer screen: participants were asked to wait while a red box was visible, and click a mouse button as quickly as possible after a green box appeared (see Figure 3). Success metrics for participants on this task are (1) clicking correctly, i.e., only when a green box appears, and (2) fast reaction time in clicking when a green box appears. As introduced in Section 3.1, this experiment consists of three rounds: first, the *task-training* round without an MR headset; second, the *benchmark* round with the MR headset; third, the *withAttack* round with the headset. Each round consists of eight levels, i.e., eight instances of a green box appearing and the user needing to click.

Color Attacks. In the context of this task, we craft visual manipulation attacks named "Color Attacks" with two types of goals: (1) **Induce an incorrect reaction** and (2) **Delay a correct reaction**. We implement a total of four attacks in the *withAttack* round. These attacks are described in Figure 4, and the participant view is shown in Figure 5.

Two of our attacks involve attempting to fool the user about the color of the real-world box, by overlaying a virtual box with the wrong color. (Note that we slightly misaligned the virtual box and made it partially transparent, so that participants could still see the actual real-world stimulus.)

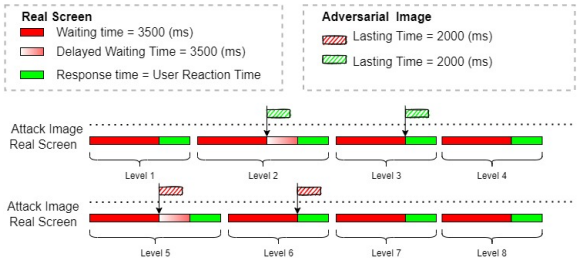
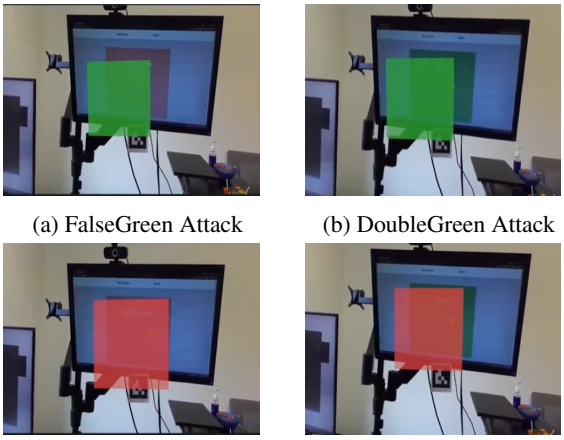


Figure 4: Timeline of attacks on the Reaction Task during the withAttack round (round 3).



(a) FalseGreen Attack (b) DoubleGreen Attack

(c) DoubleRed Attack (d) FalseRed Attack

Figure 5: Participant views of attacks in the Reaction Task — The floated boxes are generated by Color Attacks and blended into the user's view to *intentionally* mislead them.

1. **Reaction-FalseGreen** attack: Overlay a virtual green object while the actual box is still red. This attack aims to cause participants to click incorrectly.
2. **Reaction-FalseRed** attack: Overlay a virtual red object when the actual box turns green. This attack aims to mislead the participants and delay their reaction or cause them to fail to click entirely.

We crafted two other attacks, in which the virtual box color matches that of the real-world box. These attacks allow us to investigate the impact of PMA that may be startling or distracting, but is not directly misleading.

3. **Reaction-DoubleGreen** attack: Overlay a virtual green object when the actual green box appears. This attack may cause users to delay or avoid clicking as they focus on interpreting the adversarial content.
4. **Reaction-DoubleRed** attack: Overlay a virtual red object while the actual box is still red. This attack may induce users to click incorrectly, e.g., if they attempt to overcompensate for the adversarial content.

We expose participants to these attacks in the following order (see Figure 4): the FalseGreen attack on level 2, the

Table 1: Summary of all attacks.

Name	Attack Description	Attack Goal	Perception
Reaction -FalseGreen attack	Overlay a virtual green object while the real-world box is still red	Make participants click incorrectly	Visual
		Slow their reaction significantly	
Reaction -DoubleGreen attack	Overlay a virtual green object at the time the real-world box turns green	Slow their reaction significantly	Visual
Reaction -DoubleRed attack	Overlay a virtual red object while the real-world box is still red	Make participants click incorrectly	Visual
		Slow their reaction significantly	
Reaction -FalseRed attack	Overlay a virtual red object at the time the real-world box turns green	Slow their reaction significantly	Visual
Sustained Attention -NotificationSound attack	Play a sequence of notification sound during all of level 2	Make participants click incorrectly	Auditory
Sustained Attention -RingtoneSound attack	Play a sequence of ringtone sound during all of level 4	Make participants click incorrectly	Auditory
Focus -CountingCard attack	Display a sequence of playing cards	Prevent participants from noticing important context	Situational -Awareness

DoubleGreen attack on level 3, the DoubleRed attack on level 5, and the FalseRed attack on level 6. For this exploratory study, we did not randomize the order in which these attacks were presented (which would have required an infeasibly large number of participants and was not our goal).

If participants click incorrectly on level 2 (Figure 5(a), FalseGreen attack) or level 5 (Figure 5(c), DoubleRed attack), when the actual box is still red, we conclude the attack is successful in manipulating user visual perception and inducing an incorrect reaction. If the participants click correctly on level 3 (Figure 5(b), DoubleGreen attack) or level 6 (Figure 5(d), FalseRed attack), we compare their reaction time with group's average performance on the MR benchmark round. If the reaction time on valid clicks under these two attacks falls outside two standard deviations of the group's average performance, we can conclude that the attack is successful in manipulating user visual perception and slowing reaction time.

4.2 Scenario B: Sustained Attention Task, Auditory PMA

Sustained Attention Task. Sustained attention is the ability to concentrate on an activity. While naturally occurring stimuli may also disrupt users' controlled mental processing and lead to focused-attention deficit [61], we are especially interested in the capability of immersive MR audio to intentionally distract a user from another task. To evaluate the effect of auditory PMA on sustained attention, we create a scenario in which participants are asked to memorize a sequence of real-world stimuli, modeled after an existing cognitive study game [9]. For the purposes of our experiment, that real-world stimulus consists of increasingly long sequences of flashing buttons. The sequences do not build on each other but are newly random at each length. Participants are asked to memorize the sequence, and then click each button at the correct location, as shown in Figure 6. The success metric for participants on this task is to recall as many sequences correctly as possible. This experiment also consists of three rounds (*task-training*, *benchmark*, and *withAttack* rounds).

Auditory Attacks. Towards our goal of exploring different types of PMA, in this scenario we consider audio instead of visual adversarial content. Given the immersive nature of MR audio, participants might treat it as a real-world stimulus (e.g., think an actual phone is ringing). MR attackers (unlike other distractions in the user's environment) can also precisely and stealthily inject audio when participants are on high cognitive load based o n MR device sensor data.

Given the above task, we crafted two auditory attacks with one goal: (1) **Distract users at a specific point in the task**.

1. **Sustained Attention-NotificationSound** attack: Play a sequence of notification sounds during all of level 2.
2. **Sustained Attention-RingtoneSound** attack: Play a sequence of ring-tone sounds during all of level 4.

We use the *Audio Spatializer SDK* to create an immersive 3D sound effect on both audio data. The sound is played from the Oculus built-in speakers.

Figure 7 demonstrates the timeline of the *withAttack* round. If more participants fail at recalling the memorized sequence on level 2 or level 4 during *withAttack* round than *benchmark* round, we conclude that the attack is successful at affecting auditory perception and distracting participants from the primary task at the chosen times.

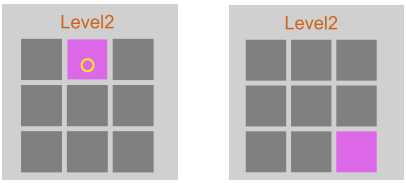


Figure 6: Two-item sequence on level 2 of the Attention Task.

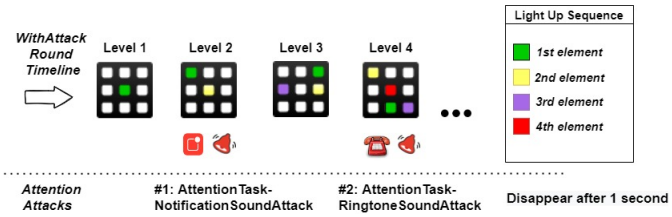


Figure 7: Timeline of the *withAttack* round (round 3) of the Attention Task. In the actual task, all elements lit up with the same color; here we use different colors to illustrate time.

4.3 Scenario C: Focus Task, Situational Awareness Attack

Focus Task. Focus is defined here as the ability to direct your attention to a particular idea [52]. As people are exploring MR usage in critical operation settings such as surgery [58], construction [44], and driving [25], researchers have raised concerns about users focusing on MR content to the detriment of other stimulus, and thus fail to notice fully-visible yet unexpected important notices. Such distraction due to MR can lead to serious danger such as falling down cliffs [13] or wandering into the street without noticing incoming vehicles [35].

To evaluate the effect of situational awareness attack to manipulate user focus, we create a scenario, parallel with the classic "Gorilla experiment" [62], in which participants' task is to notice real-world context on the monitor (though they do not know that this is their task when they begin the scenario). The text on the monitor appears while the participants are doing a decoy activity (i.e., the MR attack) with virtual content in the MR headset. The monitor text says: "If you see this message, raise your hand immediately". This real-world content becomes increasingly visible in four phases, described in the caption of Figure 8. Participants' metric for success is: notice the change in the real world as quickly as possible.

We emphasize that this task is different from the previous tasks, in that the *actual* task and success metrics are not given to participants, but rather they are given an MR activity that turns out to be the attack.

Situational Awareness Attack. Given the above task, we crafted one PMA targeting situational awareness, which refers to the perception of what is around us [22]. This attack has one goal: (1) **Prevent participants from noticing real-world instructions**. A more successful attack will keep users from noticing the real-world changes in the Focus Task for longer.

1. **Focus-CardCounting** attack: Display a sequence of playing cards to prevent participants from perceiving the real-world target.

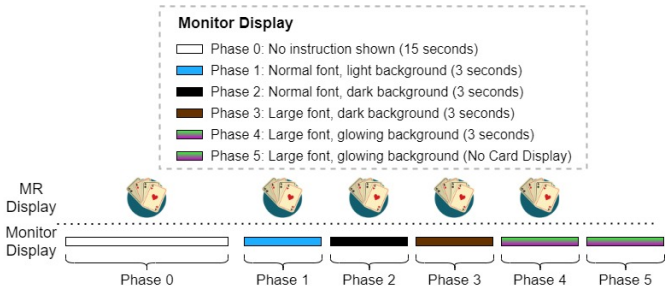
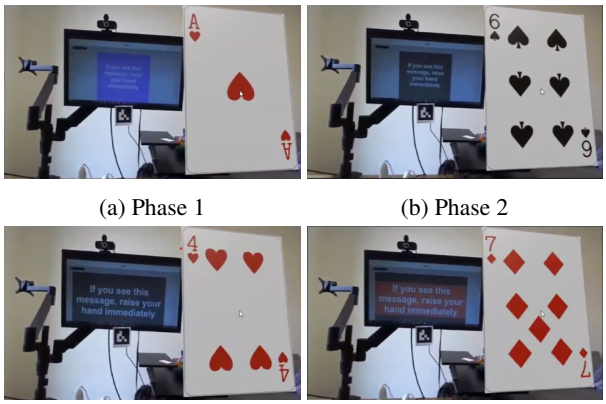


Figure 8: Timeline of the Focus Task and corresponding Situational Awareness Attack. In phase 1, the screen changes from no text displayed to our instruction. In phase 2, the background color changed from blue to black. In phase 3, the font size and the window width increased to maximum, and in phase 4, the background start to blink with different colors.



(a) Phase 1 (b) Phase 2

(c) Phase 3 (d) Phase 4

Figure 9: Participant views of the Situational Awareness Attack during the Focus Task. The floating cards are generated by the Situational Awareness Attack. Four increasingly visible phases of the real-world notification are shown on the monitor. Omitted from this figure are Phase 0 (no text on the screen) and Phase 5 (no cards in the foreground).

Unlike previous setups, this experiment only has one round, as a non-attack round would give away the nature of the task. The attack consists of 28 virtual playing cards, displayed one at a time, for one second each. Participants are asked to count the number of red cards that appeared. Starting after 15 seconds of card counting, we showed the instruction corresponding to the actual Focus Task on the computer screen. Figure 8 shows the attack and task timeline, and Figure 9 shows the participant's view. To explore user reactions under different degrees of awareness conditions, starting from the 6th participant, we advised them at the beginning of this task to stay aware of their real-world surroundings.

5 Results: Behavioral Reactions

We begin with an analysis of our experimental data to study users' reactions when encountering MR PMA (RQ1).

Impacts of Visual Attacks on Reaction Task. We use two metrics to evaluate visual PMA effectiveness: *(1) invalid click rate*, which measures the number of participants out of 21 who clicked incorrectly (before the real-world box shown on the computer monitor turns green) for a given attack, and *(2) delayed click rate*, which measures the number of participants out of 21 who took a significantly longer time to click on the real-world stimulus. Here, we define "significantly" as outside two standard deviations of the group's average performance in the previous *benchmark* round.

Table 2 summarizes the attack efficacy of all four Color Attacks. For the DoubleGreen and FalseRed attacks, recall that the real-world box turned green at the time of the attack, so Metric (1) (invalid click rate) does not apply (i.e., all clicks were valid); thus, Metric (1) is reported only for the FalseGreen and DoubleRed attacks. For the FalseGreen and DoubleRed attacks, some participants avoided the initial attack and *did* manage a valid click after the real-world box later turned green; in those cases, we evaluated their performance under Metric (2) (delayed clicks rate of valid clicks).

Figure 10 details individual participants' performance under each attack. The top of this figure shows reaction time performance for valid clicks under each attack, compared with the participant's benchmark performance. The bottom part of the figure counts that participant's invalid clicks.

Participants were susceptible to manipulative MR content that tried to evoke the target reaction (i.e., clicks). As the first row in Table 2 shows, almost all of our participants were affected by the FalseGreen attack. That is, most participants were fooled by the adversarial virtual green box and invalidly clicked in response (15/21). While our study allowed us to observe this only in our specific experimental setting, it provides a proof-of-concept demonstration, suggesting one type of user impact as a result of perceptual manipulation. This observation allows us to hypothesize that this finding would generalize to other settings where people perform tasks requiring quick reaction times while viewing MR content.

Participant reactions were slowed by manipulative MR content. As the Metric (2) (Delayed Click Rate) column in Table 2 shows, all four of our attacks delayed participants' reaction times on valid clicks. We can consider two cases here: first, attacks that occurred when the real-world box turned green, i.e., the DoubleGreen and FalseRed attacks. In both cases, the attack caused significant delays in participants responding to the real-world green box (14/21 and 16/21 participants respectively). In the second case, with the FalseGreen and DoubleRed attacks, a click at the initial time of the attack would have been invalid (a successful attack per Metric (1)); some participants avoided clicking falsely then, and made it

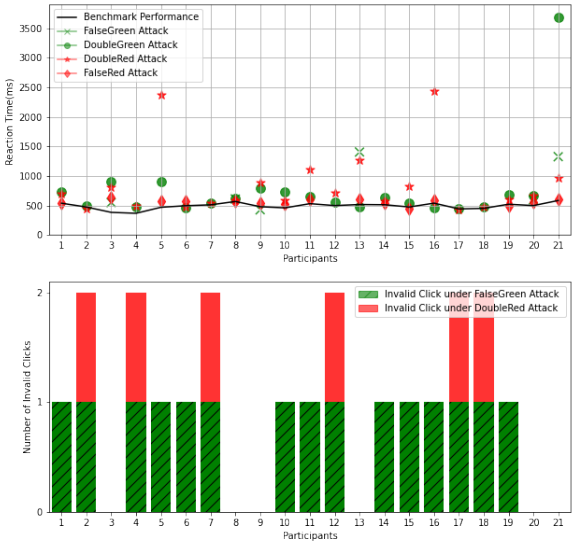


Figure 10: Per-participant performance on different Reaction Task attacks. The top graph captures every valid click, and the bottom bar chart captures the number of invalid clicks. For visual clarity we connect the benchmark dots with a line, not to imply points between participants on the line.

to the point in time when the real-world box turned green (refer to the timeline of the attacks in Figure 4). However, even in these cases, participant responses were often delayed (5/6 and 13/15 participants respectively).

We observed a few cases of extremely delayed responses — in particular, a few participants (P5, P16, P21) took over 2000 ms to click on some attacks (see Figure 10). Because the manipulative MR content was programmed to disappear after two seconds, this means that these participants waited until after the virtual box had disappeared to click. We cannot determine from our experimental results *why* participants were slowed by these attacks: whether they were distracted or confused by the manipulative MR content, whether they were attempting to avoid manipulative content, or whether they believed it was real-world content. We return to these questions in our qualitative analysis in Section 6 later.

Participant reactions were triggered by non-target manipulative MR content. We saw above that in the FalseGreen attack, participants were induced to click even though the real-world box was not green. As the third row in Table 2 shows, we also find that some (6/21) participants clicked invalidly under the DoubleRed attack, where the real-world box was also not green — even though the manipulative MR content was red. By contrast, in the *benchmark* round (with the MR headset but without any attack) in Table 2, we see that no participant *ever* clicked while the real-world box was red. In this case, we hypothesize that participants were induced to click not

Table 2: Experimental results for the Reaction Task, comparing different attacks, non-attack, and benchmark conditions. * In the benchmark round, 5% of clicks are delayed by definition (since we defined delayed clicks as those outside of 95% of the benchmark data). We give percentages for ease of interpretation, not to imply generalizability to a broader population..

Color Attacks	Metric (1): Invalid Click Rate			Metric (2): Delayed Click Rate			Combined
	Invalid Click	Total Clicks	Percentage	Delayed Click	Valid Clicks	Percentage	All
FalseGreen (round 3)	15	21	71.42%	5	6	83.33%	95.24% (20/21)
DoubleGreen (round 3)	0	21	0%	14	21	66.67%	66.67% (14/21)
DoubleRed (round 3)	6	21	28.57%	13	15	86.66%	90.48% (19/21)
FalseRed (round 3)	0	21	0%	16	21	76.19%	76.19% (16/21)
No Attack (round 3)	0	84	0%	33	84	39.29%	39.29% (33/84)
Benchmark (round 2)	0	168	0%	10*	168	5.95%*	5.95% (10/168)*

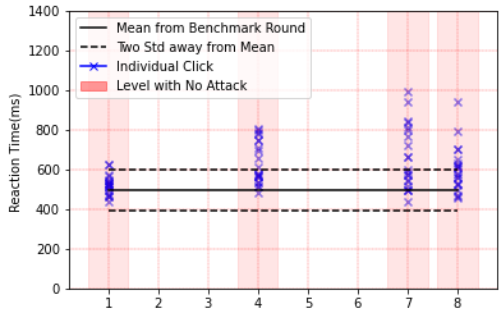


Figure 11: For the Reaction Task, comparing participants' performance on *non-attack* levels in the *withAttack* round with the group's average performance in the *benchmark* round.

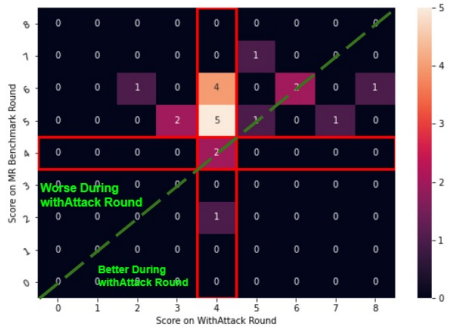


Figure 12: Results from the Attention Task. The x-axis shows scores during the *withAttack* round, and the y-axis shows scores during the benchmark round. Each box contains the number of participants who received that combination of scores. The red lines highlight that most people achieve scores of 5 or 6 in the benchmark round, while most people only reach 4 in the *withAttack* round.

because they thought the manipulative MR content was real, but because they were surprised or distracted by it, or because (having encountered some manipulative content already) they tried incorrectly to compensate for it. We again return to these questions in our qualitative analysis in Section 6 later.

Secondary impacts: reduced reaction time performance in non-attack settings. We have thus far discussed only *direct* impacts of the Color Attacks. However, we also observed an *indirect/secondary* impact. Specifically, when no attack occurred, participants still took a significantly longer time to click *after* having encountered at least one attack. Figure 11 summarizes the performance on each non-attack level during the *withAttack* round and compares it with the group's average performance during the *benchmark* round. At level 1, no attack had yet been encountered. In level 4, when participants just experienced FalseGreen and DoubleGreen attacks, we noticed a significant increase in average reaction times for nearly half of the participants. In level 7, after the FalseRed and DoubleRed attacks, we notice similar numbers of impacted participants, but with a more scattered distribution. This result suggests that even when attacks do not appear, past attacks can still impact participants' reaction process. In our setting, we saw participants become more cautious and slow down after attacks manifested, and they became even more cautious after experiencing different types of attacks.

Impacts of Auditory PMA on Sustained Attention Task. We use one metric to evaluate audio attack effectiveness, i.e., *failure rate*, which measures the number of participants out of 21 who failed at correctly recalling the provided sequence on a level when an audio attack played. Recall that we experimented with two audio attacks: the NotificationSound at level 2, and the RingtoneSound at level 4.

Manipulative MR audio content impacted participants' performance on the Sustained Attention task. We find that the NotificationSound attack on level 4 impacts many participants. The heatmap in Figure 12 shows the performance of each participant on the *benchmark* round and the *withAttack* round. Each cell in the heatmap represents the number of participants who finished at level [y] on the MR benchmark round, and at level [x] on the *withAttack* round. Overall, we find that significantly more participants (12) failed on level 4 in the *withAttack* round compared with the previous *benchmark* round (2). As above, we cannot say from our experimental results why this attack was effective, but we provide participants' self-described reflections in Section 6.

Table 3: Experimental results for the Focus Task under the Card Attack, with original and updated instructions (where participants were told to pay attention to real-world context).

	Original Instruction		Updated Instruction	
	Raise	Total	Raise	Total
Phase 1	1	5	1	16
Phase 2	0	5	4	16
Phase 3	1	5	0	16
Phase 4	0	5	1	16
Phase 5	3	5	10	16

Participants were resilient to auditory PMA under some conditions. While the NotificationSound attack on level 4 was effective, we found that the RingtoneSound attack was much less effective. Referring again to Figure 12, we see that most participants progressed past level 2 and the RingtoneSound attack during the *withAttack* round. We suspect (and some participants mentioned) that the task was sufficiently simple at earlier levels.

Impacts of Card Attack on Focus Task. We use one metric to evaluate the Card Attack's effectiveness: *phase number* (1-5), which measures when participants saw and reacted to instructions in the real world (i.e., when they raised their hands). Higher phase numbers mean that participants reacted later, that is, the attack was more effective.

Manipulative MR content prevented participants from reacting to the real-world instruction. Of our 21 participants, only two participants immediately noticed the real-world instruction and raised their hand (phase 1). In contrast, 13 participants reacted to the instruction only *after* the Card Attack was completely gone (phase 5). Despite the increasing visibility of the real-world instruction, only six participants noticed it during phases 2-4 (see Table 3).

Manipulative MR content still distracts participants even when instructed otherwise. Of the 16 participants to whom we gave updated instructions to pay attention to the real world during the scenario: only six reacted before phase 5. This suggests the unexpectedness of manipulative MR content, while sightly remedied, can still leave many participants vulnerable.

6 Results: User-Reported Reflections

From our observations of participants performing tasks and for our subsequent interview, we identified several points during our evaluation when significant human-MR interactions or tensions arose about which we gathered participant reflections (RQ2): (1) reactions to attacks, (2) mitigating the onset of attacks, and (3) making sense of the attacks. We organize our results around these points with mappings to the corresponding scenarios and attacks.

We lost the transcript for P3 because we failed to save the Zoom recording, but we did capture P3's task performance data. Here we thus report qualitative results only for 20 participants. We slightly edited some quotes for readability.

(1) Self-Reported Impacts of PMA. In Section 5, we observed attack impacts by comparing participant performance on several key metrics. Here, we turn to participants' self-reported reactions, asking them to walk us through what went through their mind when attacks occurred, helping us better understand why attacks may have been successful (or not).

Attack impact: Not knowing how to proceed . During the Reaction Time Task and Sustained Attention Task when participants first experienced PMA, many (9 of 20) were not sure what it was or how to respond. For example, P16 justified her delayed response time (over 2000 ms) when she first encountered the FalseGreen attack:

> *Because like I knew that I was supposed to follow like the color, like when it turns green on the screen. But like if it turns green outside of the screen, I was like, should I follow that? Or should I keep following the screen? (P16)*

Participants also described trying to make sense of the FalseRed attack:

> *If there is a red screen coming in front of your eyes when there is actually a green screen in the background, that's more conflicting than what I would expect it to be. (P7)*

Attack impact: Inability to focus on the primary task. Another commonly mentioned (9 of 20) impact from the Reaction Time Task and Sustained Attention Task was that the PMA distracted participants from focusing on the primary task. For example, P8 related his initial impression when experiencing one of the Color Attack:

> *It was like a lot slower because I was distracted by the green square that was popping up. (P8)*

P14 suggested the Card Attack prevented her from noticing the primary task:

> *I can't multitask, so you're like, "Pay attention to your surroundings," and I'm like, "I got to pick one, so I pick the cards. (P14)*

Attack impact: Inability to distinguish between virtual and real. Though our manipulative MR content was not close to full fidelity, and the Color Attack objects were even misaligned, we found that some participants on first impression were unable to identify the manipulative MR content as being virtual. For example, P6 later described not realizing the stimulus was virtual in Color Attack even after clicking on it multiple times:

> *I think the first color was green, and when it popped up, I clicked it, and I was like, it didn't work so I kept clicking it because it should work. (P6)*

We found that more participants (7 of 20) initially treated the audio attack as a real-world sound. For example, when P14

first heard the RingtoneSound attack, he described believing it was actually coming from a physical phone:

> *What is going on? Somebody's calling? (P14)*

Attack Impact: Entangling manipulative audio output with primary task. Six participants discussed how auditory attacks impacted their decision-making process in the Sustained Attention Task. P11 mentioned that:

> *I heard sound when I was doing the clicking and because the rhythm of the sound is different from the box changing colors, I got distracted I don't know how many times. (P11)*

P1 discussed the mental overload of handling video and audio at the same time:

> *Some of [the attacks] don't line up with what you're seeing, extra processing that you're having to do, or extra filtering to do those things... Maybe those are different parts of the brain, and those parts of the process might not really overlap. (P1)*

Though visual virtual content remains predominant in today's MR platforms and applications, our finding nevertheless suggests potential risks as MR increasingly incorporates *different* output modalities (e.g., auditory, haptic). We hypothesize that future PMA may be able to leverage multiple sensory modalities to be particularly effective.

(2) Mitigating Onset of Attacks. We now turn to participants' adaptation or defensive strategies when experiencing PMA and the subsequent impacts of their chosen strategies. Though participants did not typically interpret the attack outputs as being malicious, we can still learn from how they attempted to avoid the manipulative MR content.

Defensive technique: Mentally filtering out attack content. Many participants (8 of 20) tried to filter the attack content out of their awareness and concentrate on the non-affected area during the Reaction Time Task and Sustained Attention Task. P13 described that:

> *I think instead of noticing the [visual attack], I try to concentrate on what's behind it. (P13)*

Defensive technique: Learning from past attacks. Once participants realized that attacks were occurring (even if they did not think of them as malicious but rather as glitches), some participants adapted their behavior, anticipating and reacting more quickly to subsequent Color and Audio Attacks. For example, P6 explained:

> *At one point I kind of got used to it, and then when it flips colors, it took less time to get used to. (P6)*

Defensive technique: Physically swipe it away. We noticed that when the FalseGreen attack appeared, two participants instinctively raised their hand and tried to swat away the virtual green box. While participants did not discuss this approach, we think this natural reaction suggests potential avenues for future MR systems to detect manipulative content.

When defensive techniques fail under changing attacks. Once some participants developed a particular defensive strategy and/or adapted to a given attack, they often expected that similar attacks would occur. When the Color Attack instead changed, for example, P18 explained how he found himself newly impacted:

> *The red object tripped me up because the green object was in sync with the span. So I thought, "Oh, if I see the objects, I can just click on the feedback loop." Then when I saw the red object, I clicked on it. Obviously, I wasn't supposed to. (P18)*

Side effects from defensive techniques. Participants reported that attempting to avoid manipulative MR content was challenging. And though we found that while defensive strategies sometimes helped avoid attacks, they also caused participants to become more cautious and slower, as P1 described:

> *I think it takes a little bit more mental effort to like filter those out. (P1)*

This finding supports our experimental results from Figure 11, which showed that participants' reactions times were slowed even under non-attack conditions, after experiencing attacks.

(3) Attribution and Interpretation of PMA. Participants attributed the attacks they experienced to different causes and/or interpreted them in different ways. Before the debrief, we asked participants to share their thoughts and feelings about the experiment, and describe anything that impacted their performance. If they responded by asking about glitches, for example, we did not directly debrief with our research goal, but asked them to first elaborate on their thoughts.

Thought the attack outputs were glitches. We found that the majority of participants (14 of 20) initially assumed that the unexpected outputs in the Reaction Time Task and Sustained Attention Task were glitches, sometimes thinking back to their previous MR/VR experiences. For example, P1 with VR gaming experience recalled:

> *This is absolutely similar to some of my experiences, like when you're playing a game, and the game glitches out a little bit. (P1)*

The fact that participants often assumed (at least at first) that the attacks were glitches could in part reflect the experimental setting: we intentionally did not prime participants about security or the possible presence of attacks, and they may have given the study and the researchers the benefit of the doubt by not jumping to conclusions about malicious intentions. Still, we consider this finding to be meaningful. First, we stress that even participants who attributed the attacks to glitches were impacted by the attacks in practice. Moreover, in real MR settings, users may also be disinclined to assume the presence of malicious adversaries, and the fact that MR

software glitches are already common experiences may allow MR attacks to "hide" under the cover of such glitches.

Thought the attack output was supposed to help them. In other instances, participants assumed that the manipulative MR content in the Reaction Time Task and Sustained Attention Task was actually intended to support the primary task. For example, P8 speaks about the Color Attacks:

> *Oh it's here to like maybe help me with the task and like actually performed better. (P8)*

In the Sustained Attention Task, some participants also tried to link the audio output with the visual task. P9 assumed it was aimed to help them perform better:

> *I think I started hearing some like beeping noises... I wondered at first if maybe that was a way to give me a hint.(P9)*

This assumption was again perhaps the result of the experimental setting — expecting that the study was about testing how MR content might help someone perform a task — but we emphasize that people may make such assumptions in real MR settings, as well. Indeed, we observed that participants' trust levels towards MR were relatively high in general and that they had not previously experienced or even considered attacks in MR. Such an assumption presents a possible opportunity for attackers to either make their attacks more stealthy and/or more directly influence people's behaviors on a task.

Attributed attack output to part of the study. Ten participants noticed something was off and guessed (correctly) that it might be a deliberate part of the study, as P13 suggested:

> *I would assume if you guys are conducting the experiment, you would have it done correctly. You would stop the game if it was going awry. (P13)*

Two participants successfully identified the full purpose of the study, with one participant even spontaneously bringing up the gorilla experiment on which our Selective Attention Task was based. We stress that even these participants were still *impacted* by attacks, suggesting that suspecting attacks is not enough to protect users.

7 Discussion

While a growing body of prior work has contended with PMA in MR, our work is the first, to our knowledge, to experimentally understand end users' reactions, interpretation, and defensive strategies when experiencing PMA. We highlight several key lessons from our work, and we reflect on implications for MR designers and paint a future research vision.

7.1 Key Lessons

1. User can be manipulated by adversarial MR content even despite today's technical limitations. PMA has the ability to manipulate a user's perception of the real world (e.g., treating PMA as if it originated from the real world) and jeopardize their performance on main tasks.
2. In addition to the direct impacts of PMA, we also documented *secondary* effects that manifested on subsequent tasks or task instances — for example, participants becoming more cautious and slow on non-attack tasks after experiencing PMA.
3. Upon experiencing attacks, we observed participants developing a variety of hypotheses, including that the adversarial MR outputs were glitches, outputs were real, or outputs were supposed to help them, to explain the adversarial MR content — but participants were nevertheless impacted by them. Such expectations can be leveraged by real attackers to either make their attack more stealthy and/or more manipulative.
4. We observed cases of participants successfully adapting to the potential presence of attacks and performing better to subsequent attacks. Meanwhile, there were also examples of participants' adaptive or defensive strategies backfiring — particularly when the attack goal changed.

7.2 Implications for MR Defenses

We re-emphasize the call from prior work: MR system and application designers *must* take into account the possibility of adversarial content. Our work, along with others', experimentally demonstrates that such attacks can have real impacts on people using MR systems, and we expect that the impacts in more critical applications and/or with more finely-tuned attacks may be substantially worse. In terms of *how* to take these concerns into account, our findings with real participants position us to make several recommendations:

Contextual focus mode. As our results and previous research indicated, user can be intentionally manipulated or unintentionally distracted by MR stimuli, which can affect their performance on critical tasks. Future MR systems should take that into consideration and incorporate different levels of engagement. For example, inspired by users' defensive strategy in Section 6, when a user is on high cognitive load, future MR systems could minimize the amount of displayed information or filter other MR content out of the user's field of view.

Escape to reality. When buggy or malicious content inevitably occurs, user should be able to safely exit back to reality. Similar to the "control-alt-delete" concept for PCs, future MR systems should allow the users to easily and reliably exit the MR view when they wish, i.e., with all MR outputs verifiably disabled (as also suggested in [55]). This mechanism could also be used by users to verify whether something they perceive is MR content or part of the physical world.

Explore human-centered defenses. Though prior work has already begun to propose defenses for PMA [14, 39], one gap we observe is that these defenses fail to utilize users' reactions. Our results suggest the potential of incorporating user behavior as part of the defenses. For example, for adversarial content that aims to hold the user's attention (such as the Card Attack), a defense might involve tracking a user's

gaze and attention to virtual content, and dimming the virtual content if changes are detected in the real-world background. As another example in Section 6, we observed participants instinctively attempting to swat away adversarial visual content with their hands; MR systems could detect such reactions and remove (or offer removal) of content accordingly.

Build human resilience against attacks. We observed that some participants were able to perform better on subsequent attacks after they had been exposed to other attacks (though this was not uniformly the case). We encourage future work to further study the potential impact of prior exposure to adversarial MR content on future attack resilience, and to explore how to best take advantage of such resilience. For example, can people be trained or "inoculated" against some types of manipulative MR content through periodic exercises?

Leverage our experimental methodology to evaluate the effectiveness of proposed defenses. Besides enabling the evaluation of PMA, our methodology, which involves exposing participants to PMA in a controlled real-world environment, can be used to measure the impacts of proposed defenses above. As mentioned in Section 1, we have made publicly available our experimental testbed implementation.

Attribution of MR content. Given the rising integration of various third-party tools in the MR development cycle, we hypothesize future PMA are likely to manifest in the wild through imported third-party malicious code. Current MR systems render first-party content and third-party content in a similar format, which makes it hard for the user to distinguish if the rendered content is originated from a trusted source. We believe future MR systems should explore ways of providing trusted indicators about the source of content.

7.3 Future Directions

Anticipating future PMA in MR. Based on the interviews with participants, we speculate the possibility of even more effective PMA. For example, attacks that simultaneously combine adversarial visual and auditory outputs, or attacks that shift strategies over time to undermine users' defensive adaptations. While our work is early in the evolution of MR technologies (and hence early in the evolution of PMA), it would seem reasonable to assume that adversaries — once they manifest — may conduct their own experiments to maximize the impact of their attacks. Thus, future studies must also attempt to anticipate and protect against such threats.

Evaluating PMA in real-world settings. While we chose to conduct our experiment in a lab setting given safety concerns, a number of related works have already started to implement MR in real life scenarios such as driving [25] and walking [73]. While these works focus more on exploring the technical possibility with MR, future security researchers could apply a similar methodology to implement specialized PMA and investigate their impact on users with proper safety precautions.

Exploring PMA in a multi-user setting. Adversarial MR content might come — as in our study — from a malicious application, but it might also come from other users. As online metaverse platforms start to emerge, toxic and abusive behavior has already been observed in a multi-user settings [7], and some research has begun to explore security and privacy for multi-user AR content sharing [51, 57]. Future studies should also investigate user perception of and reaction to PMA under different multi-user dynamics.

8 Conclusion

Our goal in this work has been to explore experimentally the spectrum of end user reactions, perceptions, and defensive strategies as a result of MR-based perceptual manipulation attacks (PMA). In order to do so, we created a variety of tasks and attack scenarios and observed how users responded, adapted to, and reasoned about them. We view our contribution as laying the groundwork for continued study of PMA. To that end, our work presents a PMA evaluation framework, surfaces several key lessons from user reactions, and proposes directions for future defenses. By constructing PMA targeting different perceptions, and conducting in-depth interviews learning about user perception now, we are taking steps toward securing the full-fledged MR applications of the future.

Acknowledgments

We would like to thank our shepherd and the anonymous reviewers for their valuable feedback. We are especially grateful to our user study participants. We would also like to thank the following people for feedback on study design and/or the paper itself: Pardis Emami-Naeini, Sandy Kaplan, Kiron Lebeck, Kentrell Owens, Kimberly Ruth, Mattea Sim, Miranda Wei, and Eric Zeng. This work was supported in part by the U.S. National Science Foundation under Awards CNS-1651230 and CNS-1565252, as well as by gifts from Google, Meta, Qualcomm, and Woven Planet.

References

[1] Apple shows AR/VR headset to board. https://www.bloomberg.com/news/articles/2022-05-19/apple-shows-headset-to-board-in-sign-it-s-reached-advanced-stage.

[2] Facebook Reality Labs: Wristband for AR. https://tech.fb.com/ar-vr/2021/03/inside-facebook-reality-labs-wrist-based-interaction-for-the-next-computing-platform/.

[3] Hololens 2. https://www.microsoft.com/en-us/hololens/buy.

[4] Inside Facebook Reality Labs Research: The Future of Audio. https://about.fb.com/news/2020/09/facebook-reality-labs-research-future-of-audio/.

[5] Mark Zuckerberg demonstrating Meta's high-end Project Cambria VR headset. https://www.theverge.com/2022/5/12/23068536/meta-project-cambria-vr-ar-demo-mark-zuckerberg.

[6] Oculus Quest 2. https://www.oculus.com/quest-2/.

[7] A researcher's avatar was sexually assaulted on a metaverse platform. https://www.businessinsider.com/researcher-claims-her-avatar-was-raped-on-metas-metaverse-platform-2022-5.

[8] Rev speech-to-text services. https://www.rev.com/.

[9] A simple tool to measure your memory ability. https://humanbenchmark.com/tests/sequence.

[10] A simple tool to measure your reaction time. http://humanbenchmark.com/tests/reactiontime.

[11] Spectacles from Snapchat. https://www.spectacles.com/.

[12] Tobii XR eye tracking system. https://vr.tobii.com/oem/.

[13] Two men fall off a cliff playing Pokémon Go. https://www.latimes.com/local/lanow/la-me-ln-pokemon-go-players-stabbed-fall-off-cliff-20160714-snap-story.html.

[14] Surin Ahn, Maria Gorlatova, Parinaz Naghizadeh, Mung Chiang, and Prateek Mittal. Adaptive fog-based output security for augmented reality. In *Proceedings of the Morning Workshop on Virtual Reality and Augmented Reality Network*, 2018.

[15] Stefano Baldassi, Tadayoshi Kohno, Franziska Roesner, and Moqian Tian. Challenges and new directions in augmented reality, computer security, and neuroscience–part 1: Risks to sensation and perception. *arXiv:1806.10557*, 2018.

[16] Stuart Bender and Billy Sung. Fright, attention, and joy while killing zombies in virtual reality: A psychophysiological analysis of VR user experience. *Psychology & Marketing*, 2021.

[17] Mark Billinghurst and Hirokazu Kato. Collaborative mixed reality. In *Proceedings of the First International Symposium on Mixed Reality*, 1999.

[18] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2), 2006.

[19] Peter Casey, Ibrahim Baggili, and Ananya Yarramreddy. Immersive virtual reality attacks and the human joystick. *IEEE Transactions on Dependable and Secure Computing*, 2019.

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

[21] David Drascic and Paul Milgram. Perceptual issues in augmented reality. In *Stereoscopic Displays and Virtual Reality Systems III*, volume 2653, pages 123–134. Spie, 1996.

[22] Mica R Endsley. A taxonomy of situation awareness errors. *Human Factors in Aviation Operations*, 3(2):287–292, 1995.

[23] Elizabeth Fehrer and David Raab. Reaction time to stimuli masked by metacontrast. *Journal of Experimental Psychology*, 63(2):143, 1962.

[24] Lucas Silva Figueiredo, Benjamin Livshits, David Molnar, and Margus Veanes. Prepose: Privacy, security, and reliability for gesture-based programming. In *IEEE Symposium on Security and Privacy (SP)*, 2016.

[25] Florin-Timotei Ghiurău, Mehmet Aydın Baytaş, and Casper Wickman. ARCAR: On-Road Driving in Mixed Reality by Volvo Cars. In *Adjunct Publication of the 33rd Annual ACM Symposium on User Interface Software and Technology*, 2020.

[26] David Goedicke, Alexandra WD Bremers, Hiroshi Yasuda, and Wendy Ju. Xr-oom: Mixing virtual driving simulation with real cars and environments safely. In *13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 2021.

[27] Jeremy Raboff Gordon, Max T. Curran, John Chuang, and Coye Cheshire. Covert embodied choice: Decision-making and the limits of privacy under biometric surveillance. In *CHI Conference on Human Factors in Computing Systems*, 2021.

[28] Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology*, 74(6):1464, 1998.

[29] Jassim Happa, Mashhuda Glencross, and Anthony Steed. Cyber security threats and challenges in collaborative mixed-reality. *Frontiers in ICT*, 6:5, 2019.

[30] Harry Helson and Joseph A Steger. On the inhibitory effects of a second stimulus following the primary stimulus to react. *Journal of Experimental Psychology*, 64(3):201, 1962.

[31] Wolfgang Hoenig, Christina Milanes, Lisa Scaria, Thai Phan, Mark Bolas, and Nora Ayanian. Mixed reality for robotics. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015.

[32] Charles E Hughes, Christopher B Stapleton, Darin E Hughes, and Eileen M Smith. Mixed reality in education, entertainment, and training. *IEEE Computer Graphics and Applications*, 25(6):24–30, 2005.

[33] Suman Jana, David Molnar, Alexander Moshchuk, Alan Dunn, Benjamin Livshits, Helen J Wang, and Eyal Ofek. Enabling fine-grained permissions for augmented reality applications with recognizers. In *22nd USENIX Security Symposium*, 2013.

[34] Suman Jana, Arvind Narayanan, and Vitaly Shmatikov. A scanner darkly: Protecting user privacy from perceptual applications. In *IEEE Symposium on Security and Privacy*, 2013.

[35] Bellal Joseph and David G Armstrong. Potential perils of peri-Pokémon perambulation: the dark reality of augmented reality? *Oxford Medical Case Reports*, 2016(10), 2016.

[36] Ernst Kruijff, J Edward Swan, and Steven Feiner. Perceptual issues in augmented reality revisited. In *IEEE International Symposium on Mixed and Augmented Reality*, 2010.

[37] Eike Langbehn, Frank Steinicke, Markus Lappe, Gregory F Welch, and Gerd Bruder. In the blink of an eye: leveraging blink-induced suppression for imperceptible position and orientation redirection in virtual reality. *ACM Transactions on Graphics (TOG)*, 37(4):1–11, 2018.

[38] Joseph S Lappin and Charles W Eriksen. Use of a delayed signal to stop a visual reaction-time response. *Journal of Experimental Psychology*, 72(6):805, 1966.

[39] Kiron Lebeck, Tadayoshi Kohno, and Franziska Roesner. How to safely augment reality: Challenges and directions. In *17th International Workshop on Mobile Computing Systems and Applications*, 2016.

[40] Kiron Lebeck, Kimberly Ruth, Tadayoshi Kohno, and Franziska Roesner. Securing augmented reality output. In *IEEE Symposium on Security and Privacy (SP)*, 2017.

[41] Kiron Lebeck, Kimberly Ruth, Tadayoshi Kohno, and Franziska Roesner. Towards security and privacy for multi-user augmented reality: Foundations with end users. In *IEEE Symposium on Security and Privacy (SP)*, 2018.

[42] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proc. of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019.

[43] Paul Milgram and Fumio Kishino. A taxonomy of mixed reality visual displays. *IEICE TRANSACTIONS on Information and Systems*, 77(12):1321–1329, 1994.

[44] H Frank Moore and Masoud Gheisari. A review of virtual and mixed reality applications in construction safety literature. *Safety*, 2019.

[45] Kizashi Nakano, Daichi Horita, Nobuchika Sakata, Kiyoshi Kiyokawa, Keiji Yanai, and Takuji Narumi. Deeptaste: Augmented reality gustatory manipulation with gan-based real-time food-to-food translation. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2019.

[46] David M Neyens and Linda Ng Boyle. The effect of distractions on the crash types of teenage drivers. *Accident Analysis & Prevention*, 39(1):206–212, 2007.

[47] B Keith Payne. Prejudice and perception: the role of automatic and controlled processes in misperceiving a weapon. *Journal of personality and social psychology*, 81(2):181, 2001.

[48] Xiaolan Peng, Jin Huang, Linghan Li, Chen Gao, Hui Chen, Feng Tian, and Hongan Wang. Beyond horror and fear: Exploring player experience invoked by emotional challenge in vr games. In *Extended abstracts of the CHI Conference on Human Factors in Computing Systems*, 2019.

[49] Parinya Punpongsanon, Daisuke Iwai, and Kosuke Sato. Softar: Visually manipulating haptic softness perception in spatial augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 21(11):1279–1288, 2015.

[50] Parinya Punpongsanon, Daisuke Iwai, and Kosuke Sato. Flexeen: Visually manipulating perceived fabric bending stiffness in spatial augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 26(2):1433–1439, 2018.

[51] Shwetha Rajaram, Franziska Roesner, and Michael Nebeling. Designing privacy-informed sharing techniques for multi-user ar experiences. In *VR4Sec: 1st International Workshop on Security for XR and XR for Security*, 2021.

[52] Vilayanur S Ramachandran. *Encyclopedia of Human Behavior*. Academic Press, 2012.

[53] Derek Reilly, Mohamad Salimian, Bonnie MacKay, Niels Mathiasen, W Keith Edwards, and Juliano Franz. Secspace: Prototyping usable privacy and security for mixed reality collaborative environments. In *ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, 2014.

[54] Franziska Roesner and Tadayoshi Kohno. Security and privacy for augmented reality: Our 10-year retrospective. In *VR4Sec: 1st International Workshop on Security for XR and XR for Security*, 2021.

[55] Franziska Roesner, Tadayoshi Kohno, and David Molnar. Security and privacy for augmented reality systems. *Communications of the ACM*, 57(4):88–96, 2014.

[56] Franziska Roesner, David Molnar, Alexander Moshchuk, Tadayoshi Kohno, and Helen J Wang. World-driven access control for continuous sensing. In *ACM Conference on Computer and Communications Security*, 2014.

[57] Kimberly Ruth, Tadayoshi Kohno, and Franziska Roesner. Secure multi-user content sharing for augmented reality applications. In *28th USENIX Security Symposium*, 2019.

[58] Yu Saito, Maki Sugimoto, Satoru Imura, Yuji Morine, Tetsuya Ikemoto, Shuichi Iwahashi, Shinichiro Yamada, and Mitsuo Shimada. Intraoperative 3D hologram support with mixed reality techniques in liver surgery. *Annals of surgery*, 2020.

[59] Daisuke Sakai, Kieran Joyce, Maki Sugimoto, Natsumi Horikita, Akihiko Hiyama, Masato Sato, Aiden Devitt, and Masahiko Watanabe. Augmented, virtual and mixed reality in spinal surgery: A real-world experience. *Journal of Orthopaedic Surgery*, 28(3):2309499020952698, 2020.

[60] Susanne Schmidt, Gerd Bruder, and Frank Steinicke. Depth perception and manipulation in projection-based spatial augmented reality. *PRESENCE: Virtual and Augmented Reality*, 27(2):242–256, 2018.

[61] Richard M Shiffrin and Walter Schneider. Controlled and automatic human information processing: II. perceptual learning, automatic attending and a general theory. *Psychological Review*, 84(2):127, 1977.

[62] Daniel J Simons and Christopher F Chabris. Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, 28(9):1059–1074, 1999.

[63] Ivo Sluganovic. *Security of mixed reality systems: authenticating users, devices, and data*. PhD thesis, 2018.

[64] Qi Sun, Anjul Patney, Li-Yi Wei, Omer Shapira, Jingwan Lu, Paul Asente, Suwen Zhu, Morgan McGuire, David Luebke, and Arie Kaufman. Towards virtual reality infinite walking: dynamic saccadic redirection. *ACM Transactions on Graphics (TOG)*, 37(4):1–13, 2018.

[65] Ivan E. Sutherland. A head-mounted three-dimensional display. In *Fall Joint Computer Conference, American Federation of Information Processing Societies*, 1968.

[66] Yujie Tao, Shan-Yuan Teng, and Pedro Lopes. Altering perceived softness of real rigid objects by restricting fingerpad deformation. In *ACM Symposium on User Interface Software and Technology*, 2021.

[67] Jan Theeuwes. Exogenous and endogenous control of attention: The effect of visual onsets and offsets. *Perception & psychophysics*, 49(1):83–90, 1991.

[68] Wen-Jie Tseng, Elise Bonnail, Mark Mcgill, Mohamed Khamis, Eric Lecolinet, Samuel Huron, and Jan Gugenheimer. The dark side of perceptual manipulations in virtual reality. In *CHI Conference on Human Factors in Computing Systems*, 2022.

[69] John Vilk, David Molnar, Benjamin Livshits, Eyal Ofek, Chris Rossbach, Alexander Moshchuk, Helen J Wang, and Ran Gal. SurroundWeb: Mitigating privacy concerns in a 3D web browser. In *IEEE Symposium on Security and Privacy*, 2015.

[70] Victoria R Wagner-Greene, Amy J Wotring, Thomas Castor, Jessica Kruger, Sarah Mortemore, and Joseph A Dake. Pokémon go: Healthy or harmful? *American Journal of Public Health*, 107(1):35, 2017.

[71] Christian Weichel, Manfred Lau, David Kim, Nicolas Villar, and Hans W Gellersen. Mixfab: a mixed-reality environment for personal fabrication. In *ACM CHI Conference on Human Factors in Computing Systems*, 2014.

[72] Graham Wilson and Mark McGill. Violent video games in virtual reality: Re-evaluating the impact and rating of interactive experiences. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*, 2018.

[73] Jackie Yang, Christian Holz, Eyal Ofek, and Andrew D Wilson. Dreamwalker: Substituting real-world walking experiences with a virtual reality. In *ACM Symposium on User Interface Software and Technology*, 2019.

[74] Steven Yantis and John Jonides. Abrupt visual onsets and selective attention: evidence from visual search. *Journal of Experimental Psychology: Human Perception and Performance*.

[75] Eisa Zarepour, Mohammadreza Hosseini, Salil S Kanhere, and Arcot Sowmya. A context-based privacy preserving framework for wearable visual lifeloggers. In *IEEE PerCom Workshops*, 2016.

[76] Yi Zhang, Dan Li, Hao Wang, and Zheng-Hui Yang. Application of mixed reality based on Hololens in nuclear power engineering. In *International Symposium on Software Reliability, Industrial Safety, Cyber Security and Physical Protection for Nuclear Power Plant*, 2019.

[77] M Tarek Ibn Ziad, Amr Alanwar, Moustafa Alzantot, and Mani Srivastava. Cryptoimg: Privacy preserving processing over encrypted images. In *IEEE Conference on Communications and Network Security (CNS)*, 2016.

Appendices

Appendix A Recruitment & Screening Survey

Our primary recruiting messages were short, announcing the study and sharing a link to our recruiting survey, which provided significantly more information (see below). The recruiting messages were sent to members of our institution over Slack, mailing lists, or other private messages.

The recruiting message: "Hello everyone, I am looking for students who might be interested in participating in a user study wearing a Mixed Reality headset. The goal is to compare your performance on certain tasks with or without the MR headset. We will follow necessary COVID precautions with open windows in the user study room. The study is around 60 minutes and we will pay you $30 in Amazon gift card for your valuable time. Please let me know if you have any questions."

The full screening survey: "Thank you for taking the survey. We are a group of researchers from the University of Washington, Paul G. Allen School, and we are hoping to evaluate the impact of wearing a mixed reality headset while conducting a primary task. We appreciate your interest in our experiment and would like to conduct a quick survey beforehand. This study will take place (with COVID-19 precautions in place) on the University of Washington campus. We will reach out to you to schedule the experiment separately. This study has been reviewed by the University of Washington Human Subjects Review Board (IRB)."

1. How many times have you used a AR/MR/VR headset (HTC Vive, Oculus Rift, Windows Mixed Reality, etc.)?
   (i) I never tried it.
   (ii) I tried it a few times.
   (iii) I am a regular user.
   (iv) I use it everyday.
2. How do you feel when doing tasks in AR/MR/VR? [Open-ended]
3. Do you experience nausea when using AR/MR/VR headsets?
   (i) N/A or I'm not sure.
   (ii) No.
   (iii) Yes.
4. Do you feel eye strain when using AR/MR/VR headsets?
   (i) N/A or I'm not sure.
   (ii) No.
   (iii) Yes.
5. Do you feel dizziness when using AR/MR/VR headsets?
   (i) N/A or I'm not sure.
   (ii) No.
   (iii) Yes.
6. The headset we are using for our experiment is not compatible with some types of glasses frames. If you participate in the experiment, will you be able participate without glasses?
   (i) I don't need glasses/contacts.
   (ii) I will wear contacts to the experiment.
   (iii) I will wear glasses and if they are incompatible, I will participate without them.
   (iv) I'm not sure.

Appendix B Interview Script

Notes: As is standard with semi-structured interviews, not all interviews followed exactly this script, as researchers may have followed up on participants' responses or otherwise reordered, omitted, or adapted questions according to the context in the moment. We began each study by following COVID-19 safety procedures (e.g., sanitizing equipment).

B.1 Warm Up Phase

Thank you for participating in our research. Before we begin the study, we'd like to give you a chance to review and sign this consent form. This study has been approved by UW Human Subject Research review board. You may experience mild discomfort from using the mixed reality device or some level of motion sickness or vertigo. We will make sure that your MR headset is adjusted correctly to minimize these risks. You will also be asked to stay seated during the task, minimizing the risk of motion sickness or bumping into any real-world objects. You may choose to end the experiment at any time, without loss of promised compensation.

With your permission, we'd like to video record the study. You can still participate in the study even if you'd prefer not

to be audio or screen recorded, and you can ask us to delete the recording at any time later.

This study will have three parts: We prepare a demo app in Mixed Reality to get you used to the environment, and will ask you some follow up questions. Then we will ask you to conduct three different tasks both with and without the MR headset. At last, we will have a comprehensive discussion at the end of our experiment to learn about your experience.

1. What do you think about MR?
2. Tell us a bit about your prior MR exposure, including devices or apps that you have used or observed others using, as well as in literature or film that you have seen.
3. How do you feel about completing tasks in MR?
4. If you don't feel comfortable completing tasks in MR, what concerns do you have?

B.2 Experiment Phase

In this part of the study, we'd like you to try some games with and without the MR headset. We are not comparing your performance with others, and we focus on evaluating this technology approach. But we still hope that you try your best.

As you are completing the tasks in MR, feel free to vocalize any reactions. If you experience any buggy situation, please feel free to vocalize them as well — we won't interrupt your task to answer them, but we'd be happy to discuss them later on. Again, if you experience any severe dizziness or discomfort, feel free to end the experiment at any time.

B.3 Post-Task Interview Phase

1. How do you like the MR experience?
2. What stood out to you the most?
3. [One researcher selected one or more of the participant's experimental results to describe to them.] Is there anything that you think impacted your performance in these experiments?
4. If you were affected by the content, could you go over that moment and elaborate on it?
5. If you noticed the misleading content and successfully performed the task, could you go over that moment and elaborate on it?
6. What would you attribute the misleading content to?
7. [If participants talked about bugs and attacks] At what points do you feel the content is buggy vs the content is actually an attack?
8. What mitigating strategy did you use during the attacks?

B.4 Debrief

Now we would love to debrief with you our research purpose. The goal of our research is to evaluate whether it is possible to design mixed reality applications that mislead participants given today's technology, and measure its efficacy based on your performance. Our assumption is that, under a time or attention limited condition, people may rely on their instinct or intuition to make decisions. If the virtual generated objects blending in our physical world are similar enough to real ones, they have the ability to trigger our intuition to either make false judgement, or impact our performance.

B.5 Post Debrief Questions

- Reflect on their performance when PMA occurred.
- How has this experiment changed your trust towards AR/MR/VR?
- How will this type of technology affect our daily life in ten years?
- Do you have any concerns about adapting this technology in your daily life?

Appendix C Qualitative Codebook

The full codebook, with themes and subthemes, from qualitatively analyzing user reflection on PMA. Codes were not mutually exclusive.

Attribution of attacks

- The attack was a bug from the device or a glitch from the application.
- The attack was a part of the real world.
- The researcher deliberately programmed the attack for some reason.
- Identify the purpose of the attack and this study.

Self-reported impact of attacks

- Thought they were not impacted by the attacks.
- Thought they were impacted by the attacks because they didn't know how to proceed.
- Thought they were impacted by the attacks because they were distracted by the attacks.
- Thought they were impacted by the attacks because they believed the attack content was a part of the real world.
- Thought they were impacted by the attacks because they believed the attacks were in a different modality (audio).
- Didn't notice the attack.

Developed defensive strategies

- Focus more on the task.
- Concentrate on non-affected areas.
- Mentally filter out the attack content.
- Learn from past attacks and ignore them in later tasks.
- Try to swipe the attack content away.

User reflection on effectiveness of defensive strategies

- Thought the strategies were useful.
- Thought the strategies made them more cautious and thus react slower.
- Thought the strategies were not sufficient, and thus user was still affected by attacks.
- Thought the strategies they developed for one attack backfired against another attack.