Exploring the Privacy Risks of Adversarial VR Game Design

Vivek Nair*
UC Berkeley
Berkeley, CA, USA
vcn@berkeley.edu

Gonzalo M. Garrido*
TU Munich
Munich, Germany
gon.munillag@gmail.com

Dawn Song
UC Berkeley
Berkeley, CA, USA
dawnsong@berkeley.edu

James F. O'Brien
UC Berkeley
Berkeley, CA, USA
job@berkeley.edu

ABSTRACT

Fifty study participants playtested an innocent-looking "escape room" game in virtual reality (VR). Within just a few minutes, an adversarial program had accurately inferred over 25 of their personal data attributes, from anthropometrics like height and wingspan to demographics like age and gender. As notoriously data-hungry companies become increasingly involved in VR development, this experimental scenario may soon represent a typical VR user experience. Since the Cambridge Analytica scandal of 2018, adversarially-designed gamified elements have been known to constitute a significant privacy threat in conventional social platforms. In this work, we present a case study of how metaverse environments can similarly be adversarially constructed to covertly infer dozens of personal data attributes from seemingly-anonymous users. While existing VR privacy research largely focuses on passive observation, we argue that because individuals subconsciously reveal personal information via their motion in response to specific stimuli, active attacks pose an outsized risk in VR environments.

KEYWORDS

virtual reality, metaverse, data harvesting, privacy, anonymity

1 INTRODUCTION

Through the fog of rapidly shifting preferences for internet platforms, one clear trend has stood the test of time: with each new and improved medium for accessing the web comes a new and improved method for harvesting personal user data. As these technologies become more immersive and tightly integrated with our daily lives, so too do the corresponding intrusive attacks on user privacy.

Privacy concerns have existed since the early days of the world wide web, even when users primarily accessed information through static websites. The emergence of social media platforms in the early 2000s changed this paradigm, generating a torrent of additional data on user behavior. Third-party (tracking) cookies that can uniquely identify and follow individuals [11] around the web then allowed this data to be deployed for everything from surveillance advertisement [12] to pushing political agendas [49].

In the past decade, users shifted to accessing the web primarily via their mobile phones (92.1% as of 2022 [64]), simultaneously introducing a suite of newly-extractable data attributes like audio, video, and geolocation. Next, the wave of wearable devices such

as smart watches added critically sensitive data like biometrics and health information into the mix [67]. Most recently, virtual home assistants have made possible pernicious intrusions into users' most private activities [18]. Overall, the tendency is clear: each new technology has gradually expanded the scope of data attributes accessible to attackers, further eroding any expectation of privacy.

Virtual reality (VR) is well-positioned to become a natural continuation of this trend. While VR devices have been around in some form since well before the internet [52], the true ambition of major corporations to turn these devices into massively-connected social "metaverse" platforms has only recently come to light [42, 60, 65]. These platforms, by their very nature, turn every gaze, movement, and utterance of a user into a stream of data, broadcast to other users around the world in the name of facilitating real-time interaction.

It has long been understood that individuals exhibit distinct biomechanical motion patterns that can be used to identify them or infer their personal attributes [13, 27, 31, 33, 48, 53], which researchers have shown can be exploited to identify and profile users in VR [23, 35, 41, 44]. While existing work has largely focused on passive observation of VR users, the success of games specifically designed to harvest user data [25] on conventional social platforms motivates us to investigate similar active attacks in VR.

Gaming has been the predominant driver of the recent uptick in VR adoption, with 91 of the 100 most popular VR applications being games [66]. This paper aims to shed light on the significant privacy risks associated with adversarially-designed VR games that appear innocuous to end users. We have identified over 25 examples of private data attributes that attackers can covertly harvest from VR users, which we experimentally demonstrate in our 50-person user study. Many of these attributes would be difficult to observe passively but can be obtained with high fidelity by prompting users to unknowingly reveal more information about themselves via carefully designed interactive game elements. We aim to increase broad awareness of privacy concerns within VR and encourage privacy practitioners to examine the challenges and solutions that lie at the intersection of data privacy and the VR-enhanced social internet.

The main contributions of our study are:

(1) We provide a comprehensive framework of virtual reality information flow, threat actors, data sources, observable attribute classes (§2), and systematic privacy attacks (§3).

(2) With our open-source VR demo, we illustrate how malicious game developers could design adversarial yet seemingly-innocuous VR environments that trick users into revealing personal information through their motion and behavior patterns (§4).

(3) We experimentally demonstrate how an attacker can covertly harvest over 25 unique data attributes from VR users, many of which are difficult to obtain through passive observation (§5).

Attacker Type	Data Sources					Observable Attribute Classes				
	Raw Sensor Data	Processed Telemetry	Rendering Pipeline & Host System APIs	Networked Telemetry	Presented Telemetry	Device	Network	Geospatial	Audio	Behavior
Privileged Attacker I	✓	✓	✓			✓		✓	✓	✓
Privileged Attacker II		✓	✓	✓		✓	✓	✓*	✓	✓
Privileged Attacker III				✓	✓		✓	✓*	✓*	✓
Non-Privileged Attacker					✓			✓*	✓*	✓

*Observable only in weaker filtered/preprocessed format
Table 1: Virtual reality threat actor capabilities.

2 BACKGROUND

The aim of this section is to provide background information on virtual reality devices and applications necessary to understand the substance of this paper. We outline a holistic framework of VR information flow and threat actors by which the current contributions can be clearly differentiated from the existing literature.

2.1 VR Information Flow

A VR device uses an array of external or onboard sensors to generate a stream of information about its user. In a typical consumer-grade VR system, the position and orientation of a head-mounted display (HMD) and two hand-held controllers are measured in 3D space (six degrees of freedom per tracked object) at a rate of between 60 Hz and 144 Hz, resulting in a "telemetry stream."

Users can download various games and applications for their VR device from an app store provided by the manufacturer. These applications consume the telemetry stream generated by the VR device and use it to render stimuli for the user, thereby creating an immersive experience. In the case of virtual telepresence (or "metaverse") applications, the telemetry data is also shared with a variety of external systems. The typical information flow for such an application, as depicted in Fig. 1, is as follows:

(1) The VR device processes raw sensor data into useful telemetry, which it provides to the application via an API. The application uses this data to provide stimuli to the user via a rendering pipeline, which the application completely controls.
(2) If the application involves interactions with other users, the client-side VR application streams processed telemetry data to an external server via a network to facilitate such interactions.
(3) The server then relays this data to other users so that an "avatar" of the original user can be rendered on their devices.
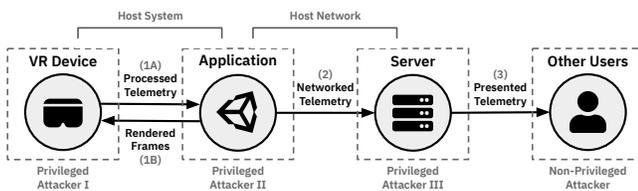

Figure 1: Virtual reality information flow and threat model.

The general basis of the VR privacy threat is that telemetry information available to an adversary can be used to infer private user data instead of being used for its intended purpose. Accordingly, our VR threat model considers four potential adversaries corresponding to the four distinct entities associated with the VR information flow. We summarize the capabilities of each attacker in Table 1.

2.2 VR Threat Model

Privileged Attacker I (the "Hardware Adversary").
The first privileged attacker represents the party controlling the firmware of a target user's VR device. This attacker has access to raw sensor data from the VR device, including spatial telemetry, audio/visual streams, and device specifications. There is a bi-directional information flow between the device and the local application: the device provides processed telemetry to a running application, which the attacker can manipulate arbitrarily, and the device receives a stream of audio/visual stimuli from the application, which the attacker can manipulate arbitrarily before presenting to the user. However, this attacker cannot read or manipulate the network communications of the application.

Privileged Attacker II (the "Client Adversary").
Our second privileged attacker represents the developer of the client-side VR application running on the target user's device. This attacker has full access to the APIs provided by the VR device and host system and controls a graphics rendering pipeline that the attacker can use to provide visual stimuli to the target user. In the case of a multiplayer or server-based application, the attacker can process this data arbitrarily before streaming it to a server.

Privileged Attacker III (the "Server Adversary").
Our third privileged attacker represents the entity controlling the external server used to facilitate multiplayer functionality for the application running on the target user's device. This entity may, in practice, be the same party developing the client-side application (in the case of a "public server") or an entirely separate entity (a "private server"). Thus, the same entity often controls Privileged Attackers II and III. This attacker receives a stream of telemetry data from the client-side application, which it can process arbitrarily before relaying such data to one or more other client devices. The inferences available to this attacker are generally weaker than the previous attackers; for example, a client application may receive tracking data at 120 Hz and broadcast it at 30 Hz [73], and audio signals are typically heavily compressed before being broadcast.

Non-Privileged Attacker (the "User Adversary").
A non-privileged attacker represents a second end-user of the same multiplayer application as the target user. This attacker receives low-fidelity telemetry and audio streams from the external server for legitimate purposes, such as rendering an avatar representing the target user. They can also interact with the target user as permitted by the application, such as to provide stimuli and observe the target's response. While the audio and telemetry streams are likely highly processed and filtered by this point, they are typically still sufficient to observe the general behavior of the target user.

The goal of this paper is to understand the capabilities of an adversarial VR application. As such, we evaluate all threats from the perspective of Privileged Attacker II, while noting which other adversaries may be capable of performing the same attacks.

2.3 Observable Attribute Classes

We now shift our discussion to the broad classes of private user data observable by each of the attackers using only their corresponding data sources. We categorize the collected attributes into primary (captured directly from a data source), secondary (derived deterministically from primary attributes), and inferred (derived from primary and secondary attributes using machine learning).

Spatial Telemetry. The first major source of user data is directly from telemetry (namely, the position and orientation of the VR headset and controllers over time). Such data is useful for revealing certain anthropometric measurements, such as height and wingspan. While all attackers can observe telemetry to some extent, less privileged attackers are likely to experience degraded precision when estimating these metrics due to the use of intermediate filtering and processing. For example, we found that privileged attackers I and II can determine a user's interpupillary distance (IPD) from telemetry to within 0.1mm, but IPD is difficult for privileged attackers III and non-privileged attackers to ascertain.

Device Specifications. Another class of attack aims to use VR-specific heuristics to determine information about the VR device and the user's host computer. Of course, privileged attackers I and II can directly query device specifications such as resolution and field of view (FOV) from available system APIs; however, we will later demonstrate how even non-privileged attackers can attempt to learn some of this information by creating puzzles that only users of high-fidelity devices can feasibly solve. Determining the specifications of a user's device can reveal personal information about the users themselves. For instance, the cost of commercially-available VR setups spans at least two orders of magnitude; determining the exact hardware of a target user may reveal their income/wealth.

Network Observations. An additional source of information about a target user is the observation of network characteristics. While not necessarily unique to virtual reality, attacks that leverage network observations to geolocate users are a natural fit for virtual telepresence applications, which often facilitate the use of multiple game servers to minimize perceived latency [73]. Thus, privileged attackers II and III can efficiently execute such attacks.

Behavioral Observations. Behavioral observations are a fourth key source of private information enabled by virtual reality applications, and observing how users react to carefully chosen stimuli can reveal a wide variety of personal information. Attacks based on observing user behavior typically require less privilege than other types of attacks discussed herein, with even non-privileged attackers typically receiving enough information to observe general user interactions. We also include listening to user vocalizations in this category (audio), although one could also consider it a category.

2.4 Related Work

2.4.1 Motion. For decades, researchers have shown that individuals reveal information about themselves via their motion. As early as 1977, Cutting and Kozlowski demonstrated that people can identify their friends just by viewing the motion of 8 tracked points affixed to the body [13], and that the gender of the participants could be identified by a stranger with statistically significant accuracy [33]. More recently, Pollick et al. (2005) [53] used statistical techniques to achieve accurate identification of gender from motion, and Jain et al. (2016) [27] found that the age of a participant can also be accurately determined by their motion patterns.

2.4.2 VR Privacy. Because virtual reality devices generate a stream of data relating to the motion of a user, and people are known to subconsciously reveal information about themselves via their motion, it is natural to question the extent to which personal attributes are inferable in VR. According to popular literature reviews of the field [10, 14, 16, 19, 23, 24, 34, 39, 42, 50, 61, 70], the vast majority of existing VR privacy research has focused on passive observation.

One major area of interest for researchers has been uniquely identifying users based on their observed motion patterns in VR. A variety of laboratory studies [36, 51, 51, 72] have been conducted in which 16 to 511 users were passively observed while performing a number of standardized tasks in VR, after which machine learning techniques were used to uniquely identify them with accuracies ranging from 40% to 95%. More recently, Nair et al. [44] passively observed 55,541 players of a popular VR game, and were later able to uniquely identify players with 94% accuracy based on their motion.

A second line of work has focused on identifying specific demographic attributes like age and gender based on VR recordings [5, 15, 71]. As with the identification studies, these works utilize the passive observation of people using standard VR applications.

While the notion of malicious VR application design has been qualitatively discussed since 2020 [4], no user study has yet been produced that directly implements and evaluates an adversarially-designed VR application. However, the success of gamified data harvesting on traditional social media platforms suggests that even stronger capabilities may be available to an active attacker.

2.4.3 Gamified Data Harvesting. In 2018, the British political consulting firm Cambridge Analytica was revealed to be in possession of personal data from up to 87 million Facebook users. Subsequent analysis revealed that most of this data was collected through Facebook quizzes designed to seem like fun personality assessments while actually building a detailed profile of user data [8, 26, 62].

Gamified data collection mechanisms bypass the normal cognitive filters associated with data privacy by taking advantage of users' innate desire to perform optimally when completing challenges. Presenting a data-revealing question as a puzzle or element serving a legitimate role in a broader game has proven effective at obscuring the hidden intent to collect personal information [2, 3].

2.5 Motivation

In this paper, we seek to combine the known possibility of inferring private data attributes from VR motion data with the success of gamified data harvesting in conventional social platforms to explore data harvesting attacks made possible by adversarial game design in VR. Our interest in this field is motivated by two simple observations. First, gaming is the predominant driver of VR adoption today

[66], providing ample opportunity to disguise data collection mechanisms as VR game elements. Simultaneously, the social platforms exploited by Cambridge Analytica are now dominant players in the AR/VR space. It is natural to assume that some may wish to use the same techniques that have proven successful on conventional social media platforms in the data-rich environment of VR.

3 VR PRIVACY ATTACKS

Next, we describe specific examples of adversarial game elements designed to extract VR user data, corresponding to the broad observable attribute classes detailed in §2.3. The goal of this section is not to be exhaustive with respect to the wide variety of interactive elements that can reveal user information, but rather to exemplify strategies for collecting various types of attributes using specific mechanisms that we later evaluate in our user study (§4).

3.1 Biometrics
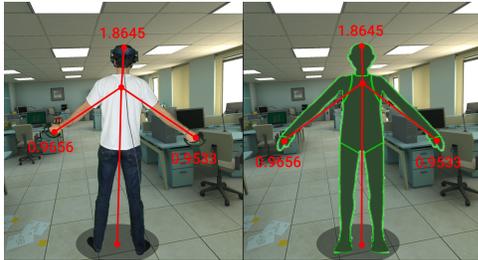
Figure 2: Measuring user anthropometrics from telemetry.

Continuous Anthropometrics. Basic anthropometrics provide a simple yet compelling example of the dangers of adversarial design. Fig. 2 illustrates how attackers can passively measure a user's height and wingspan from VR telemetry. However, users are unlikely to naturally stand in a position that readily facilitates the measurement of wingspan. Therefore, Fig. 3 depicts a pose-based game element designed to subtly induce a standing position more conducive to precise anthropometric measurement.

Figure 3: Adversarial measurement of wingspan.

Binary Anthropometrics. An attacker can collect binary anthropometrics, which include characteristics such as longer-arm and dominant handedness, both directly from telemetry (e.g., "which hand moves more?") and from behavior (e.g., "which hand is used to press a button?"). Fig. 4 illustrates an example process of determining a user's handedness by including a small button that requires precise manipulation, suggesting the use of one's dominant hand.

Figure 4: Estimating handedness from behavior.

Figure 5: VR puzzle revealing deuteranopia.

Vision. VR attackers can carefully construct interactive elements that secretly reveal aspects of a player's visual acuity, such as nearsightedness, farsightedness, or color blindness. For example, Fig. 5 shows a puzzle element of a VR game that appears innocuous to most users but is not solvable by users with red-green color blindness (deuteranopia), thus revealing the presence of that condition.
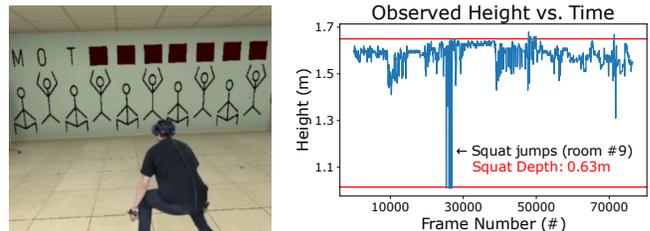

Figure 6: Measurement of physical fitness.

Fitness. Attackers can also use behavioral and telemetric measurements to asses a subject's degree of physical fitness. Fig. 6 illustrates a virtual room designed to elicit physical activity and shows the resulting metric of physical fitness measurable on a headset position (y-coordinate) vs. time graph. We observed that a squat depth of less than 25% of height corresponded to low physical fitness, though other metrics can also be used. An extreme lack of fitness may reveal a participant's age or the presence of physical disabilities.

Figure 7: VR puzzle measuring reaction time.

Reaction Time. Fig. 7 shows a VR environment adversarially constructed to reveal the participant's reaction time by measuring the time interval between a visual stimulus and motor response. Reaction time is strongly correlated with age [77].

3.2 Environment
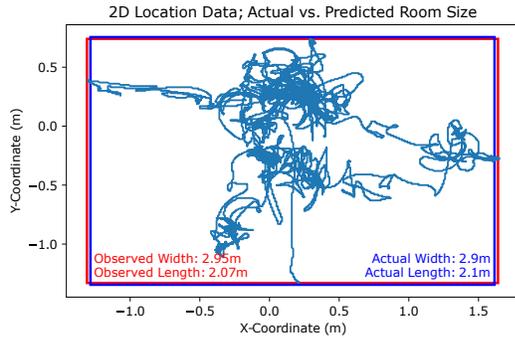
Figure 8: Estimating room size from spatial telemetry.

Room Size. Fig. 8 shows how an attacker could estimate the size of a user's physical environment by tracking their virtual movements. Virtual environments can be designed to contain interactive elements which specifically encourage the participant to explore the boundaries of their physical environment.
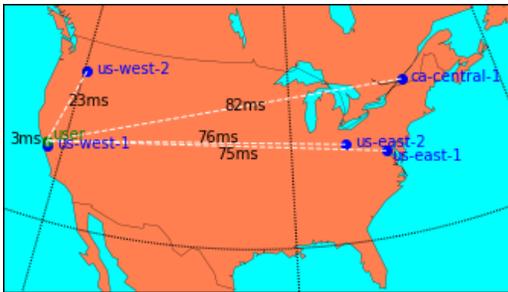

Figure 9: Estimating user location from network latency.

Geolocation. Fig. 9 shows how observing the round-trip delay between a client device and multiple game servers (proximity) can reveal an end user's location (locality) via multilateration. A non-privileged attacker could use the round trip delay of audio signals as an approximate measure of latency.

3.3 Device Specifications
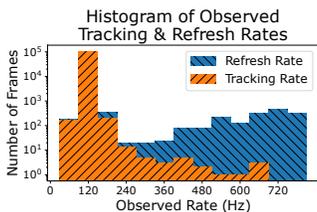
(a) Measuring headset and tracking refresh rate from device API throughput (log scale).

(b) Environment designed to reveal headset refresh rate via perceived motion differences.

Figure 10: Methods of attaining VR device metrics.

VR Device. We assume that privileged attackers I and II have intrinsic knowledge of the VR device specifications via direct API interaction. Fig. 10a shows how privileged attackers may use the observed update frequency of telemetry data to determine the polling rate of a target user's controller tracking. Further, Fig. 10b shows how even a non-privileged attacker can construct a virtual environment that replicates the "UFO test" [1], which users perceive differently depending on their devices' refresh rate (see puzzle 15 in Appendix A). Currently, determining refresh rate, resolution, and field of view is sufficient to reveal the exact model of the VR device.

Host Device. Privileged attackers can also embed a variety of standardized benchmarks in their source code to assess the quality of the target user's host device (gaming computer). An attacker can use metrics such as CPU power, GPU power, and network bandwidth to reveal the age and price tier of the system and, thus, potentially correlate the spending power of the target user.


(a) abstraction   **(b) attention**

(c) naming   **(d) orientation**
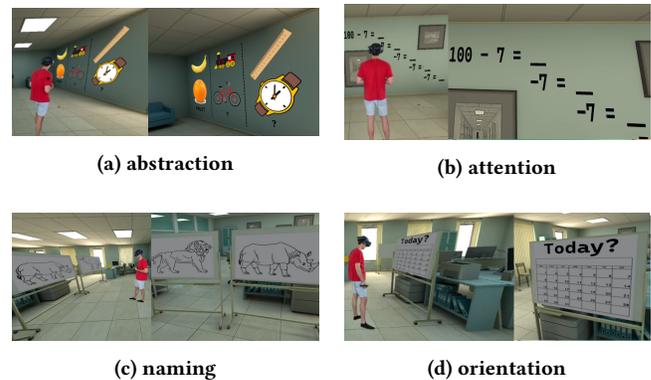
Figure 11: Methods of measuring cognitive acuity.

3.4 Acuity (MoCA)

A number of standardized cognitive, diagnostic, and aptitude tests can be adapted for (and hidden within) VR environments. Fig. 11 illustrates VR environments designed to covertly asses four categories of the Montreal Cognitive Assessment (MoCA): abstraction (11a), attention (11b), naming (11c), and orientation (11d).

3.5 Demographics

Figure 12: Determining language from user behavior.

Language. There are a number of ways to ascertain a user's spoken language(s) in VR, including via speech recognition. Fig. 12 illustrates how a non-privileged attacker can observe a user's direction of gaze while solving a puzzle to reveal the languages they speak.

	Head			Left			Right			Device	Microphone	Behavior
	X	Y	Z	X	Y	Z	X	Y	Z			
Height		✓										
Left Arm	✓		✓	✓		✓						
Right Arm	✓		✓				✓		✓			
Longer Arm	✓		✓	✓		✓	✓		✓			
Handedness				✓	✓	✓	✓	✓	✓			
Wingspan				✓		✓	✓		✓			
Room Length	✓											
Room Width		✓										
Room Size	✓		✓									
IPD										✓		
Eyesight												✓
Color Blindness												✓
Locality											✓	
Device Refresh Rate												✓
Tracking Refresh Rate										✓		
Device Resolution										✓		
Device FOV										✓		
VR Device										✓		✓
Computing Power										✓		
Languages												✓
Physical Fitness		✓										
Reaction Time												✓
MOCA												✓
Gender		✓		✓		✓	✓		✓	✓	✓	
Age		✓										✓
Ethnicity		✓									✓	✓
Disability Status (Mental)												✓
Disability Status (Physical)		✓										
Identity	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 2: VR device sensors associated with each attack.

Vocal Characteristics. Listening to the voice of a user may reveal key demographic attributes such as age, gender, and ethnicity [7, 22]. Shared VR environments with voice streaming provide a strong opportunity to exploit voice analysis, as attackers can cue target users to speak certain words or phrases that reveal more information, such as by requiring players to speak passwords aloud.

Inferred Attributes. While most demographic attributes cannot be observed directly from VR data, attackers can often accurately infer them from primary data attributes. For example, height, wingspan, and IPD correlate strongly with gender, while eyesight, reaction time, and fitness correlate with age. While not possible to measure accurately in this study, we also suggest that in practice, information about room size, VR device type, and computing power could be used together to infer the income or wealth of a user.

Identity. Finally, the various attributes discussed above can be combined to uniquely identify VR users. While previous work has already demonstrated the deanonymization of VR users [41] within small groups, the number of data attributes presented here far exceed the 15 or so thought necessary to uniquely identify every individual in the United States [58].

3.6 Summary

Table 2 summarizes the VR privacy attacks presented in this section along with the VR device sensors or sources of information associated with each attack. The incredible volume of information exposed by a metaverse user, with at least 18 telemetry values collected 60 times per second or more, provides a vast amount of data from which adversarial inferences can be made. In all, we have identified dozens of unique data attributes, ranging from biometrics and demographics to behavioral and environmental measurements, that can be observed from users in VR via adversarial game design.

Of course, these attacks are by no means exhaustive, with many further attributes likely being observable that we have not discussed. Instead, our examples serve to illustrate the wide scope of observations available to VR adversaries and the ability to capture a comprehensive user attribute profile that would otherwise have involved aggregating data across several different devices.
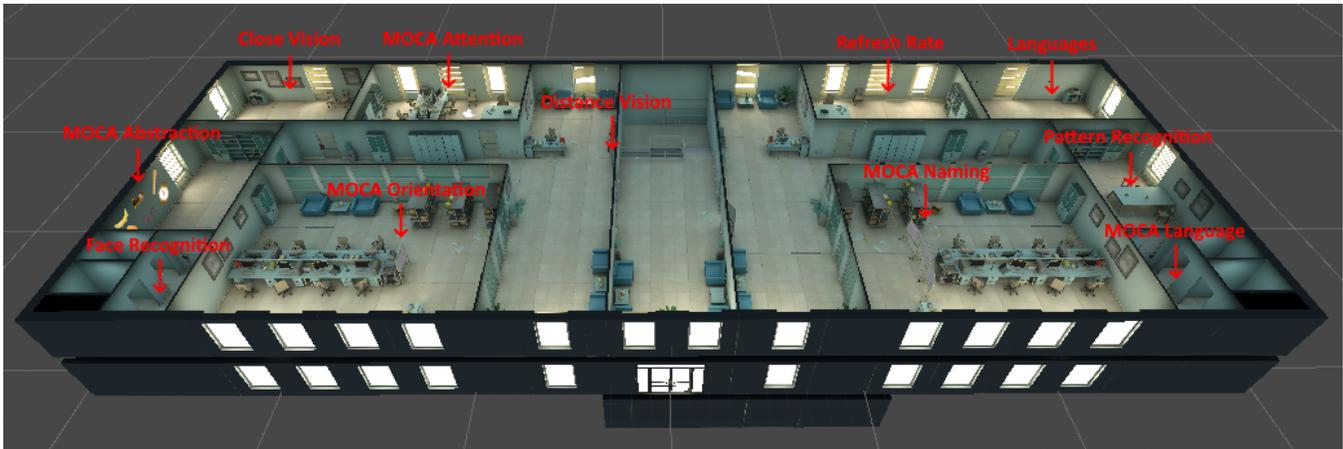
Figure 13: Virtual office building of the "MetaData" game, hosting the puzzle rooms.

Having discussed in great detail the theoretical information flow, adversaries, and plausible inferences of metaverse environments, the remainder of this paper focuses on the experimental validation and quantification of these threats.

While a detailed description of the evaluated attacks is necessary for the completeness of this study, it is not our intention to focus on any particular attributes. Rather, our goal, as highlighted by the experimental design described below, is to generally demonstrate the extent to which adversarial game design can enable the collection of sensitive data attributes in VR.

4 EXPERIMENTAL DESIGN

In this section, we describe "MetaData," a virtual reality "escape room" game designed as a case study for understanding how adversarial game design enhances attacker capabilities in virtual reality. The question we aim to answer is whether, and to what degree, an attacker can use data collected from consumer-grade VR devices to accurately extract and infer users' private information when aided by the capability to adversarially construct the virtual world and application rather than merely relying on passive observation.

This section details the experimental design, technical setup, and protocol used to answer this important question. After identifying the privacy-sensitive variables we believed to be accessible within VR (as detailed in §3), we implemented systematic methods to collect and analyze these variables from within VR applications.

To test the efficacy of these attacks, we designed an "escape room"-style VR game themed as an office building (see Fig. 13). We then disguised the attacks as a set of puzzles within the game, which users were highly motivated to solve to the best of their ability in order to unlock a sequence of doors and win the game. We describe and illustrate the exact puzzles in detail in Appendix A.

We endeavored to design the experiment such that it did not bluntly reveal the ulterior goal, thereby illustrating how other VR applications could also accomplish the same goal covertly. To this end, we also added innocuous (i.e., "noisy") rooms which did not necessarily collect meaningful personal information, but instead served to camouflage the data-harvesting puzzles.
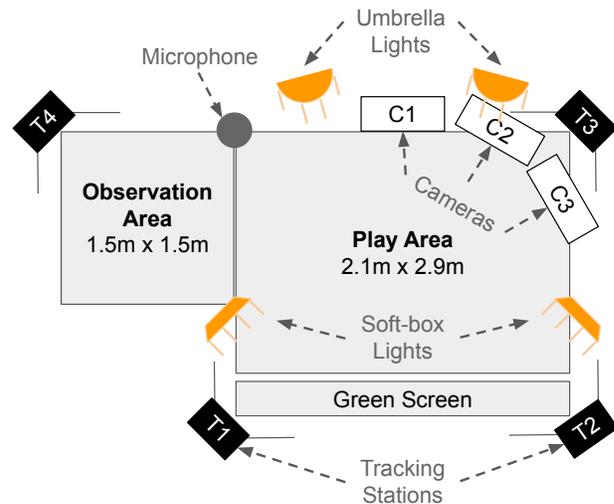
Figure 14: VR laboratory room layout.

4.1 Setup and Protocol

We recruited 50 individuals for the experiments (participant distribution given in Appendix B). After completing a thorough informed consent and orientation process, we helped the participants don a VR headset (HTC Vive, Vive Pro 2, or Oculus Quest 2) and its hand-held controllers (Vive Controllers, Valve Index Controllers, or Oculus Quest Controllers, respectively), after which the participant proceeded to play the VR game (see the laboratory room layout in Fig. 14 and the primary VR setup in Fig. 15). Finally, the participants completed a post-game survey to collect the "ground truth" values for attributes of interest. The methods for collecting the true attribute values are summarized in Appendix C.

We tested three devices to determine if there were any noteworthy differences in the findings, which we did not observe other than in IPD (see §5.1), and to provide distinct classes for device identification. Each headset was paired with a gaming computer sufficiently powerful to run it at full fidelity; the main experimental

setup had 64 GB of RAM, an AMD Ryzen 9 5950X CPU, and an Nvidia RTX 3090 GPU. To produce accurate results for room size and geolocation, we also conducted our experiment across four geographically distinct laboratories.

Each experiment lasted approximately 10–20 minutes within VR, plus around 10 minutes for completing the survey. Throughout the experiments, we minimized the interactions with the participants and ensured their safety by intervening when they approached a wall in the room. The experiments remained the same for all participants; we did not alter the game play-through or logic. The game collected the targeted data points in CSV format during the play-through. Furthermore, the researchers manually annotated data points for data collection that required game development beyond what is reasonable for this study, e.g., automating voice recognition to register the escape room "passwords" (solutions) the participants articulated aloud. The researchers pressed keys on a keyboard to trigger animations in the virtual environment and teleport the player between rooms. These elements could be automated in a production-ready VR game.

Figure 15: Experimental setup (not a real participant).

Once the experiment ended, the participants filled out a form with their ground truth, which we used to validate the accuracy of the proposed privacy attacks. To collect the ground truth unknown to the participants themselves, we performed onsite measurements, e.g., we annotated the VR device and VR-room area, tested their reaction time with a desktop app, and measured their height and wingspan with a metric tape. Furthermore, knowing that researchers have studied the use of cognitive assessments in the diagnosis of attention disorders [57], autism [28], PTSD [38],

and dementia [75], we chose the Montreal cognitive assessment (MoCA) [29] as a simple example of what advanced, immersive VR games could hide in their play-throughs. We randomized the order of the VR experiment and paper MoCA test (with half the participants taking the MoCA before and with the other half after the experiment) to neutralize potential biases in either direction. The exact method of collecting "ground truth" measurements for each attribute value is described in Appendix C.

Once we collected the ground truth, we ran our analysis scripts (privacy attacks) over the collected data to compile and infer data points, which we compared to the ground truth to assess the attacks' accuracy. The results of these experiments are described in §5 and summarized in Table 3.

4.2 Ethical considerations

We identified three primary ethical risks in our protocol: (i) the risk of discomfort using a VR device, (ii) the risk of a confidentiality breach of participant data, and (iii) the risk that participants might not have wished to disclose certain information about themselves during the course of the study.

To address the first risk (i), we used high-fidelity VR devices and appropriately powerful gaming computers for all participants, together capable of consistently providing 120 frames per second, well above the minimum specifications recommended to mitigate the risk of VR sickness [63]. We designed our VR game to avoid distressing elements such as horror, claustrophobia, or flickering/strobing lights. Furthermore, a researcher was present to ensure participants did not collide with real-world objects during each play-through.

To address the second risk (ii), we anonymized all collected data using random codes that we could not reasonably trace back to a participant's identity. Moreover, we avoided collecting any highly-sensitive data that could potentially damage participants in a breach. Lastly, we normalized biometric measurements on a scale of 0 to 1 to avoid revealing exact measurements in this paper (e.g., in Fig. 16). The photos included in this paper are not of actual participants.

To address the third risk (iii), we made sure participants clearly understood the nature of the study. We emphasize that this is not a deception study. Our claims about the non-obviousness of the presented attacks should not be construed to imply that participants were unaware that their data was being collected during the study. Participants were informed that their data was being collected, including a description of the categories of data being observed. After completing the VR portion of the study, participants were made aware of the exact attributes being collected. They were explicitly given the opportunity to withdraw consent without penalty at any point in the process, including after having detailed knowledge of the data attributes involved, in which case their data would not have been included in the results.

In light of these considerations, the study was deemed a minimal-risk behavioral intervention and was granted an IRB exempt certification under 45 C.F.R. § 46.104(d)(3) by an OHRP-registered institutional review board.

Attribute	Type / Source	Precision	Accuracy	Statistics	Attackers
Height	Primary Telemetry	1 cm	76% within 5 cm 94% within 7 cm	$R^2 = 0.75$	Privileged I-III Non-Privileged*
Longer Arm	Primary Telemetry	boolean	58% for $\geq$ 2 cm difference 100% for $\geq$ 3 cm difference	$F_1 = 0.67$ $F_1 = 1.00$	Privileged I-III Non-Privileged*
Interpupillary Distance	Primary Telemetry	0.1 mm	96% within 0.5 mm (Vive Pro 2) 58% within 0.5 mm (All Devices)	$R^2 = 0.99$ $R^2 = 0.58$	Privileged I-II
Wingspan	Secondary Telemetry	1 cm	78% within 7 cm 98% within 12 cm	$R^2 = 0.68$	Privileged I-III Non-Privileged*
Room Size	Secondary Telemetry	1 m^2	70% within 2 m^2 96% within 3 m^2	$R^2 = 0.97$	Privileged I-III Non-Privileged*
Geolocation	Primary Network	100 km	50% within 400 km 94% within 500 km	N/A	Privileged II-III
HMD Refresh Rate	Primary Device	1 Hz	100% within 3 Hz (Privileged Attacker) 88% wtihin 60 Hz (Unprivileged Attacker)	$R^2 = 0.99$ $R^2 = 0.75$	Privileged I-II Privileged III* Non-Privileged*
Controller Tracking Rate	Primary Device	1 Hz	100% within 2.5 Hz	$R^2 = 0.99$	Privileged I-II Privileged III* Non-Privileged*
Device Resolution (MP)	Primary Device	0.1 MP	100% within 0.1 MP	$R^2 = 1.00$	Privileged I-II
Device FOV	Primary Device	10°	100% within 10°	$R^2 = 0.92$	Privileged I-II Privileged III* Non-Privileged*
Computational Power	Primary Device	0.1 GHz 10 Mh/s	CPU: 100% within 0.4 GHz GPU: 100% within 20 Mh/s	$R^2 = 0.92$ $R^2 = 0.81$	Privileged I-II
VR Device	Secondary Device	categorical	100%	$p = 0.00$	Privileged I-III Non-Privileged*
Handedness	Primary Behavior	boolean	96%	$F_1 = 0.98$	Privileged I-III Non-Privileged
Eyesight	Primary Behavior	boolean	72% (Hyperopia) 80% (Myopia)	$F_1 = 0.73$ $F_1 = 0.75$	Privileged I-III Non-Privileged
Color Blindness	Primary Behavior	boolean	100%	$F_1 = 1.00$	Privileged I-III Non-Privileged
Languages	Primary Behavior	boolean	90%	$p = 0.08$	Privileged I-III Non-Privileged
Physical Fitness	Primary Behavior	boolean	86%	$F_1 = 0.92$	Privileged I-III Non-Privileged
Reaction Time	Primary Behavior	categorical	88%	$F_1 = 0.90$	Privileged I-II Privileged III* Non-Privileged*
Acuity (MoCA)	Primary Behavior	1 point	94% within 2 points 100% diagnostic accuracy	$F_1 = 1.00$	Privileged I-III Non-Privileged
Gender	Inferred Classification	boolean	98%	$F_1 = 0.98$	Privileged I-III Non-Privileged
Age	Inferred Regression	1 yr	100% within 1.5 yr	$R^2 = 0.99$	Privileged I-III Non-Privileged
Ethnicity	Inferred Classification	categorical	98%	$p = 0.01$	Privileged I-III Non-Privileged
Disability Status	Inferred Classification	boolean	100%	$F_1 = 1.00$	Privileged I-III Non-Privileged
Identity	Inferred Classification	categorical	100%	$p = 0.00$	Privileged I-III Non-Privileged

* With degraded accuracy.

Table 3: Selected attributes collected and analyzed during the experiment, with accuracy and R^2, F_1, or p values from χ^2 tests.
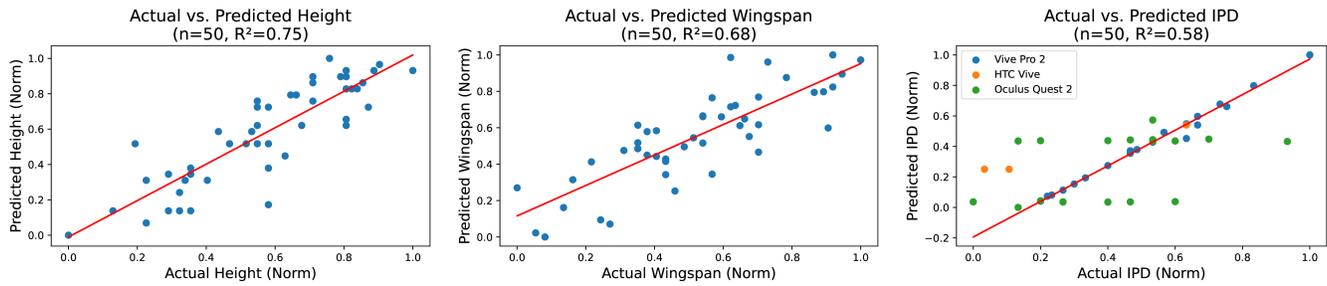
Figure 16: Actual and predicted user anthropometrics.

5 RESULTS

In this section, we present the empirical effectiveness of the privacy attacks introduced in §3, as summarized in Table 3.

5.1 Biometrics

Continuous Anthropometrics. Fig. 16 shows (scaled) actual and predicted values for *height* ($R^2 = 0.75$), *wingspan* ($R^2 = 0.68$), and *interpupillary distance* (IPD) ($R^2 = 0.58$). IPD measurements were most accurate on the Vive Pro 2, with $R^2 = 0.99$ when excluding other devices. In general, we could accurately determine these three metrics for most users from just a few seconds of telemetry. We were not, however, able to accurately predict the individual lengths of the left and right arms ($R^2 = 0.02$ and $R^2 = 0.01$ respectively), due to the lack of a reliable center point from which to measure.

Binary Anthropometrics. Although absolute arm lengths were not discernible, relative lengths were accurate enough that we could usually identify which of the participant's arms was longer. We observed increasing accuracy for participants with greater differences in length, reaching 100% accuracy for the 12% of participants with a difference of at least 3 cm. Handedness can also be determined accurately from behavioral observations; we note, however, that 94% of our participants reported being right-handed.

Vision. Our vision tests achieved diagnostic accuracies for *hyperopia* (farsightedness), *myopia* (nearsightedness), and *deuteranopia* (red-green color blindness) of 72%, 80%, and 100% respectively. The overall accuracy of detecting a visual deficiency was 80%, in part because some users of contact lenses could not remove their contacts for the experiment.

Fitness. Using squat depth as a correlate of *physical fitness* discriminated "low" fitness with an accuracy of 86%; our tests were not able to differentiate between "moderate" and "high" fitness.

Reaction Time. We measured *reaction time* to a precision of one recorded frame (16.6 ms). We were able to detect whether a participant's reaction time was above or below 250 ms (the approximate median reaction time) with an accuracy of 88%.

5.2 Environment

Room Size. The *length* and *width* of each of three testing rooms was determined to within 1.0 m with accuracies of nearly 90%. This

allowed true room area to be found within 3 m² in 96% of trials. Taking the average estimated area for each tested room vs. the true accessible room area yields $R^2 = 0.97$.

Geolocation. Using the server latency multilateration (hyperbolic positioning) technique for *geolocation* yielded a mean longitudinal and latitudinal error of around 2.5° across four tested locations. This was sufficient to locate the test subject to within 500 km in 94% of cases, and within the correct state in 100% of cases.

5.3 Device Specifications

Tracking Rate. We found that privileged attackers could determine various VR device specifications (namely, *display refresh rate*, *display resolution*, *field of view*, and *tracking rate*) with high accuracy. Tracking rate (the number of unique telemetry measurements taken per second) is a particularly interesting metric, as the top four VR headsets, together accounting for over 75% market share [20], all have different default HMD refresh rates (72/144/80/90 Hz).

VR Device. Using the above device specifications and the highly heterogeneous nature of VR device specifications, privileged attackers can determine the type of VR device with 100% accuracy. We also found that non-privileged attackers could determine the refresh rate to within 30 Hz with an accuracy of 30% and to within 60 Hz with an accuracy of 88%; however, this was not sufficient to accurately determine the type of device.

Host Device. We found that an attacker benchmarking host device specifications can determine *GPU power* with 100% accuracy to within 20 Mh/s (daggerhashimoto) and *CPU clock speed* to within 0.4 GHz, allowing them to estimate the price tier of the host device.

5.4 Acuity (MoCA)

Table 4 summarizes the numerical (continuous, i.e., the score of each category) and diagnostic (binary, i.e., passing or failing a category) accuracy of the *Montreal Cognitive Assessment* (MoCA) we conducted in the VR experiments. We achieved a diagnostic accuracy of 90% or greater for 5 of the 7 scored MoCA categories (excluding visuospatial/executive and delayed recall), with an overall diagnostic accuracy of 100%.

MoCA Category	Accuracy (Numerical)	Accuracy (Diagnostic)
Executive	N/A	N/A
Naming	100%	100%
Memory	78%	84%
Serial 7	90%	100%
Attention	88%	100%
Repetition	74%	96%
Language	74%	96%
Abstraction	100%	100%
Recall	60%	90%
Orientation	100%	100%
Overall	80% within 1 point 94% within 2 points	100%

Table 4: Accuracy of each MoCA category.

5.5 Demographics

Language. The visual focus method of *language* determination identified a spoken language (other than English) with at least conversational proficiency in 90% of multilingual participants.

Vocal Characteristics. We used existing machine learning models to determine the gender [7] and ethnicity [22] of participants from their voice with an accuracy of 98% and 66% respectively; these accuracy values improved to 100% when combined with other attributes such as height and wingspan as described in "Inferred Attributes" below.

Inferred Attributes. We used Azure Automated Machine Learning [40] to determine the optimal preprocessor, model type, and input metrics for inferring several demographic attributes. Table 5 summarizes the results of this meta-analysis. For identity, we used the best-performing technique of Miller et al. [41]. Using the identified optimal models, we determined the participant's gender, ethnicity, disability status, age (within 1.5 years), and identity with nearly 100% accuracy across several Monte Carlo cross-validations; thus, users were never simultaneously present in the training and testing datasets, other than for inferring identity, and it is not possible that the demographic inferences were a result of identifying users.

With respect to disability, we there was one reported physical disability and three reported mental disabilities amongst our 50 participants; we were able to identify these disabilities individually with 100% accuracy ($F_1 = 1.00$). In each case, the model far outperformed any individual attribute; for example, ethnicity was 98% accurate despite its most significant input (voice) being only 66% accurate on its own.

Attribute (Prediction)	Inputs	Preprocessing / Model
Gender (Classification)	Voice, Height, Wingspan, Interpupillary Distance (IPD)	TruncatedSVDWrapper SVM
Age (Regression)	Close Vision, Reaction Time, Height, Test Duration, Acuity	MaxAbsScaler ExtremeRandomTrees
Ethnicity (Classification)	Voice, Language, Height	StandardScalerWrapper LightGBM
Disabilities (Classification)	Vision, Fitness, Acuity	MaxAbsScaler NaiveBayes
Identity (Classification)	Height, Wingspan, Acuity, IPD, Vision, Reaction Time	StandardScalerWrapper RandomForest

Table 5: Inputs and methodology of inferred attributes.

By following the Azure Automated ML Workflow [40], we avoided the biases of manually selecting features for demographic inference.

The AutoML workflow begins by clustering users into training, validation, and testing sets. Using the validation set, a variety of preprocessing techniques are first evaluated. Next, a variety of classical machine learning models are trained using the preprocessed features, using the validation set to evaluate the accuracy of each algorithm. For the most accurate architecture, a hyperparameter sweep is performed to optimize the model for the feature set.

Using the best-performing model, an explainability analysis is conducted to select the most important input features for inferring each attribute. For example, rather than instructing the model to infer gender from voice, height, wingspan, etc., we initially provided the system with all available primary attributes, and allowed it to determine on its own (e.g., via PCA) which are relevant to a given inference task. Finally, the testing set is used to evaluate the selected approach, consisting of automatically-determined features, preprocessing, architecture, and hyperparameters. The process is repeated across several Monte Carlo cross-validations

6 DISCUSSION

We now return to the question of whether an attacker can use data collected from consumer-grade VR devices to extract and infer users' private information accurately. In this study, we have shown that this is indeed possible, with moderate to high accuracy values for most of the aggregated and inferred data points presented in Table 3. Using the technique of Miller et al. [41], we found that an attacker could consistently identify a participant among the pool of 50 within a few minutes of gameplay. Moreover, we have collected more than 25 granular data points, well above the 15 necessary to uniquely identify every individual in the United States [58]. While we were required to condense this data collection into a concise 20-minute experiment for logistical reasons, real-world attackers could gain increased accuracy and covertness by aggregating data collected over much longer periods of time.

6.1 Participant Awareness

In sections 3 and 4, we argued that a developer could design VR environments and games to facilitate the covert collection of targeted data points disguised as normal game elements.

For ethical reasons, our participants were informed that they were participating in an adversarial experiment where their personal information would be collected. However, they were not told exactly which attributes were being collected until the end of the experiment. Thus, our experiment reflects a realistic scenario in which users may be generally aware of privacy threats but would not necessarily know the attributes collected by a given application.

Upon debriefing, we asked the participants whether they could identify which of the attributes we were attempting to measure. All 50 participants reported not knowing exactly which attributes were being collected and inferred during the game, with no participant able to correctly identify more than 3 of the attributes.

When we revealed the list of attributes, many participants expressed surprise at the breadth of information that could be collected within VR, but none expressed particular shock at the existence of some degree of data harvesting (perhaps having already

grown accustomed to these practices on the web). Even if participants were aware of the attributes being measured, we believe it would have been difficult to counteract many of the attacks, due to the deeply subconscious nature of many of the observed behaviors.

6.2 Adversarial Capabilities

In §3.1, we provided wingspan as an initial example of an attribute that is far easier to observe in an adversarial application; while users rarely adopt a posture in which their arms are completely outstretched, an adversarial game can easily drive users to adopt such a posture through the use of an interactive element.

Reflecting now on the entire set of attributes inferred in this study, we observe that nearly all of them are aided by the introduction of adversarial puzzles in our escape room. For example, most behavioral observations, including the entire MoCA acuity assessment, relied on observing the user's responses to specific adversarially-constructed puzzles.

Our major conclusion from this finding is that the privacy risk of VR devices stems not only from their sensors, such as accelerometers and gyroscopes, but also from the immersive nature of their displays, which can be used to totally control a user's virtual environment to influence the information they reveal. Thus, while past research has primarily focused on the capability to infer various attributes from motion-tracking data, it is equally important to consider the threat posed by adversarial VR application design.

6.3 Societal Implications

While for ethical reasons we limited our attacks to relatively benign data points, an attacker could potentially track and infer additional information about other more critically sensitive personality traits, like sexual, religious, or political orientation, educational level, and illnesses, among others, to enhance practices such as surveillance advertisement [12] or pushing political agendas [49]. Given how immersive and emotionally engaging VR environments can be [30, 37, 47, 74], such practices could become more pernicious and effective than with current mobile and desktop applications.

The risk of profiling users in VR is further exacerbated by the known ability to uniquely identify users based on their VR motion data [45]. This already presents a significant privacy risk, as typical VR usage today includes rowdy gaming sessions and adult experiences that users may not want to be linkable to their identity in more professional settings. Further, by combining our profiling techniques with cross-application identification methods, an attacker, or even a group of colluding attackers, could attempt to aggregate user profiles from data across many applications.

Although we use the terms "attack" and "attacker" throughout this paper, to the best of our knowledge, there is nothing illegal about the methods described herein. It is possible that in the future, many VR users would knowingly or unknowingly consent to this form of data collection via clauses contained in platform terms of service or end-user license agreements. In fact, major VR device manufacturers have been observed selling headsets at a loss [54] of up to $10 billion per year [55], evidently with the aim to recoup said losses with some form of after-sales revenue.

6.4 Limitations

The results of this study should be understood in light of a few limitations. Unfortunately, our sample of participants was not perfectly representative of the general population; for example, college students were overrepresented. For logistical reasons, we were unable to tamper with VR device firmware and thus could not consider hardware-level attacks in this paper. Therefore, privileged attackers I and II, while different in theory, had identical capabilities within the scope of our experiment. While ground truth values for user attributes were measured by the researchers when possible, many were also self-reported (see §C) and thus potentially biased. Lastly, the researchers were forced to interact with participants outside of VR on some occasions, such as to warn of nearby obstacles. While we did attempt to minimize such occurrences, these interactions could nevertheless have biased certain results.

6.5 Future Work

Given the early stage of research on privacy and VR, there are many outstanding questions in this field for researchers to tackle. Among them is the question of how developers can design VR games or applications that make privacy attacks even more stealthy, including by integrating these attacks into daily tasks in future VR/AR environments. On the other hand, researchers could also study analysis techniques for revealing hidden data collection mechanisms (where possible) to make these attacks harder to achieve. Furthermore, studying what additional data attributes an attacker could leverage from data sources we did not consider (including eye tracking and full-body tracking) will expand our overall awareness of VR-related vulnerabilities. Additionally, future work dedicated to how an attacker could not just observe but actually change users' opinions will shed light on the implications of future immersive and impactful metaverse applications.

Above all, we believe it is most important to study the potential countermeasures to these VR privacy attacks. Unfortunately, it is difficult to defend against attacks based on inferring private attributes from VR motion data. In VR, motion data is often required for legitimate purposes, such as passive authentication or multiplayer functionality. Thus, it is difficult for a user to ascertain whether their data is being used for malicious or benign purposes without auditing the server-side logic. Nevertheless, we can propose a few broad approaches for combatting this threat.

Local Differential Privacy. In "MetaGuard" [43], researchers have demonstrated that differentially private noise sampling functions, such as the bounded Laplace mechanism, are reasonably effective at defeating VR privacy attacks. While this approach incurs a usability penalty, it does so according to a theoretically optimal trade-off between data privacy and perceived tracking error.

Adversarial Models. Outside of VR, researchers have used generative adversarial networks (GAN) to obscure sensitive data from biometric sources like gait [56]. While this approach loses the provable properties of differential privacy, thes models could actually be more effective at protecting privacy in practice, due to their ability to detect and obscure not only primary sensitive attributes but also hidden correlations to these variables. For example, if instructed to

hide a user's height, the model would quickly learn to also obscure their wingspan, which is highly correlated to the former.

Behavioral Modification. "Homuncular flexibility" refers to the innate ability of users to operate control schemes that are very different from their own bodies. For example, VR users can effectively control avatars with 10 times their normal arm length, avatars with 3 arms, or even avatars representing completely different creatures, such as crabs [76]. In doing so, they may be less likely to introduce subconscious movement patterns that reveal personal information, than when controlling an anatomically familiar avatar. Thus, homuncular flexibility could be leveraged to induce users to behave in ways that limit identifiability and profiling.

Trusted Execution Environments. Finally, there will likely remain some scenarios in which no amount of modification to tracking data is permissible. For example, in a competitive VR e-sports title, any perturbation of telemetry data could upset the scoring mechanism of the game. In such situations, trusted execution environments, such as Intel's SGX, may provide a hardware-based attestation mechanism that allows users to verify the software running on a remote machine before sending their data to that server. Thus, without modifying their motion data, users can ensure that only legitimate and necessary operations are being performed.

7 CONCLUSION

In this study, we shed light on the serious privacy risks of the metaverse by showing how VR can be turned against its users. Specifically, we provided a comprehensive security and privacy framework for VR environments that classifies (i) attackers, (ii) data sources, (iii) vulnerable attributes, and their corresponding (iv) attacks. We demonstrated the practicality and accuracy of these attacks by designing and conducting experiments with 50 participants using consumer-grade VR devices. The participants played our "escape room" VR game, which was secretly designed to collect personal information, like biometrics, demographics, and VR device and network details, among numerous other data points. The results demonstrate high information leakage with moderate to high accuracy values over most identified vulnerable attributes, with just a handful of these attributes being sufficient to uniquely identify a user [6, 17, 21, 32, 46, 59, 68, 69].

The alarming accuracy and covertness of these attacks and the push of data-hungry companies towards metaverse technologies indicate that data collection and inference practices in VR environments will soon become more pervasive in our daily lives. Furthermore, the breadth of possible VR applications, increasing quality of VR devices, and relative simplicity of our demonstration, all suggest that more sophisticated attacks with a higher success rate are possible and perhaps on the horizon. Therefore, we hope our work encourages other privacy practitioners to advance research at the intersection of privacy and VR, in particular to propose countermeasures for new and existing privacy attacks in the metaverse.

AVAILABILITY

The Unity (C#) source code and compiled binaries for the "escape room" VR game we designed for our experiments are available for download from our anonymized repository. The repository also contains the data collection instruments and data analysis scripts used to determine the primary attributes and sample data.

https://github.com/MetaDataStudyAnonymized/MetaDataStudy

ACKNOWLEDGMENTS

REFERENCES

[1] Blur Busters TestUFO Motion Tests. Benchmark for monitors & displays.
[2] Data Collection Through Gamification.
[3] HR Magazine - Why Cambridge Analytica's techniques could kill gamification, April 2018.
[4] Nadisha-Marie Aliman and Leon Kester. Malicious design in aivr, falsehood and cybersecurity-oriented immersive defenses. In *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 130–137, 2020.
[5] Abdullah Al Arafat, Zhishan Guo, and Amro Awad. Vr-spy: A side-channel attack on virtual key-logging in vr headsets. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pages 564–572, 2021.
[6] Maryam Archie, Sophie Gershon, Abigail Katcoff, and Aileen Zeng. Who's watching? de-anonymization of netflix reviews using amazon reviews, 2018.
[7] Kory Becker. primaryobjects/voice-gender, May 2022. original-date: 2016-06-09T14:30:44Z.
[8] Hal Berghel. Malice domestic: The cambridge analytica dystopia. *Computer*, 51(05):84–89, 2018.
[9] Blur Busters. UFO Motion Tests. https://www.testufo.com/. Online; accessed 30 April 2022.
[10] Kent Bye, Diane Hosfelt, Sam Chase, Matt Miesnieks, and Taylor Beck. The ethical and privacy implications of mixed reality. In *ACM SIGGRAPH 2019 Panels*, SIGGRAPH '19, New York, NY, USA, 2019. Association for Computing Machinery.
[11] Aaron Cahn, Scott Alfeld, Paul Barford, and S. Muthukrishnan. An empirical study of web cookies. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 891–901, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee.
[12] Matthew Crain. *Profit Over Privacy.* Minneapolis: University of Minnesota Press, 2021.
[13] James E. Cutting and Lynn T. Kozlowski. Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the Psychonomic Society*, 9(5):353–356, May 1977.
[14] Jaybie A. De Guzman, Kanchana Thilakarathna, and Aruna Seneviratne. Security and privacy approaches in mixed reality: A literature survey.
[15] Jaybie Agullo de Guzman, Kanchana Thilakarathna, and Aruna Seneviratne. Safemr: Privacy-aware visual information protection for mobile mixed reality. In *2019 IEEE 44th Conference on Local Computer Networks (LCN)*, pages 254–257, 2019.
[16] Ellysse Dick. Balancing user privacy and innovation in augmented and virtual reality.
[17] Cynthia Dwork, Adam Smith, Thomas Steinke, and Jonathan Ullman. Exposed! a survey of attacks on private data. Publisher: Annual Reviews.
[18] Jide S. Edu, Jose M. Such, and Guillermo Suarez-Tangil. Smart Home Personal Assistants: A Security and Privacy Review. *ACM Computing Surveys*, 53(6):1–36, November 2021. arXiv:1903.05593 [cs].
[19] Ben Falchuk, Shoshana Loeb, and Ralph Neff. The social metaverse: Battle for privacy. *IEEE Technology and Society Magazine*, 37(2):52–61, 2018.
[20] Jamie Feltham. Valve Index Is Now The Second Most Used Headset On Steam, October 2021. Section: VR Hardware.
[21] Xianyi Gao, Bernhard Firner, Shridatt Sugrim, Victor Kaiser-Pendergrast, Yulong Yang, and Janne Lindqvist. Elastic pathing: your speed is enough to track you. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and*

Ubiquitous Computing - UbiComp '14 Adjunct, pages 975–986, Seattle, Washington, 2014. ACM Press.

[22] Yatharth Garg. Speech-Accent-Recognition, May 2022. original-date: 2018-06-21T07:55:52Z.

[23] Gonzalo Munilla Garrido, Vivek Nair, and Dawn Song. Sok: Data privacy in virtual reality, 2023.

[24] Alberto Giaretta. Security and privacy in virtual reality – a literature survey, 2022.

[25] Jim Isaak and Mina J. Hanna. User data privacy: Facebook, cambridge analytica, and privacy protection. *Computer*, 51(8):56–59, 2018.

[26] Jim Isaak and Mina J Hanna. User data privacy: Facebook, cambridge analytica, and privacy protection. *Computer*, 51(8):56–59, 2018.

[27] Eakta Jain, Lisa Anthony, Aishat Aloba, Amanda Castonguay, Isabella Cuba, Alex Shaw, and Julia Woodward. Is the Motion of a Child Perceivably Different from the Motion of an Adult? *ACM Transactions on Applied Perception*, 13(4):1–17, July 2016.

[28] W. Jarrold, P. Mundy, M. Gwaltney, J. Bailenson, N. Hatt, N. McIntyre, K. Kim, M. Solomon, S Novotny, and L. Swain. Social attention in a virtual public speaking task in higher functioning children with autism. *Autism Res.*, 2013.

[29] Parunyou Julayanont and Ziad S. Nasreddine. Montreal cognitive assessment (MoCA): Concept and clinical review. In A. J. Larner, editor, *Cognitive Screening Instruments*, pages 139–195. Springer International Publishing, 2017.

[30] Orin S Kerr. Criminal law in virtual worlds, 2008.

[31] Adam G. Kirk, James F. O'Brien, and David A. Forsyth. Skeletal parameter estimation from optical motion capture data. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2005*, pages 782–788, June 2005.

[32] Daniel Kondor, Behrooz Hashemian, Yves-Alexandre de Montjoye, and Carlo Ratti. Towards Matching User Mobility Traces in Large-Scale Datasets. *IEEE Transactions on Big Data*, 6(4):714–726, December 2020.

[33] Lynn T. Kozlowski and James E. Cutting. Recognizing the sex of a walker from a dynamic point-light display. *Perception & Psychophysics*, 21(6):575–580, November 1977.

[34] Ronald Leenes. Privacy in the metaverse. In Simone Fischer-Hübner, Penny Duquenoy, Albin Zuccato, and Leonardo Martucci, editors, *The Future of Identity in the Information Society*, pages 95–112, Boston, MA, 2008. Springer US.

[35] Jonathan Liebers, Mark Abdelaziz, Lukas Mecke, Alia Saad, Jonas Auda, Uwe Gruenefeld, Florian Alt, and Stefan Schneegass. Understanding user identification in virtual reality through behavioral biometrics and the effect of body normalization. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[36] Jonathan Liebers, Mark Abdelaziz, Lukas Mecke, Alia Saad, Jonas Auda, Uwe Gruenefeld, Florian Alt, and Stefan Schneegass. Understanding User Identification in Virtual Reality Through Behavioral Biometrics and the Effect of Body Normalization. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–11, Yokohama Japan, May 2021. ACM.

[37] Junsu Lim, Hyeonggeun Yun, Auejin Ham, and Sunjun Kim. Mine yourself!: A role-playing privacy tutorial in virtual reality environment. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, CHI EA '22, New York, NY, USA, 2022. Association for Computing Machinery.

[38] L. Loucks, C. Yasinski, SD. Norrholm, J. Maples-Keller, L. Post, L. Zwiebach, D. Fiorillo, M. Goodlin, T. Jovanovic, AA. Rizzo, and BO. Rothbaum. You can do that?!: Feasibility of virtual reality exposure therapy in the treatment of ptsd due to military sexual trauma. *Anxiety Disord.*, 2019.

[39] Divine Maloney, Samaneh Zamanifard, and Guo Freeman. Anonymity vs. familiarity: Self-disclosure and privacy in social virtual reality. In *26th ACM Symposium on Virtual Reality Software and Technology*, VRST '20, New York, NY, USA, 2020. Association for Computing Machinery.

[40] Microsoft. Azure Automated Machine Learning - AutoML.

[41] Mark Roman Miller, Fernanda Herrera, Hanseul Jun, James A. Landay, and Jeremy N. Bailenson. Personal identifiability of user tracking data during observation of 360-degree VR video. *Scientific Reports*, 10(1):17404, October 2020. Number: 1 Publisher: Nature Publishing Group.

[42] Stylianos Mystakidis. Metaverse.

[43] Vivek Nair, Gonzalo Munilla Garrido, and Dawn Song. Going incognito in the metaverse, 2022.

[44] Vivek Nair, Wenbo Guo, Justus Mattern, Rui Wang, James F. O'Brien, Louis Rosenberg, and Dawn Song. Unique identification of 50,000+ virtual reality users from head & hand motion data, 2023.

[45] Vivek Nair, Wenbo Guo, Justus Mattern, Rui Wang, James F. O'Brien, Louis Rosenberg, and Dawn Song. Unique identification of 50,000+ virtual reality users from head & hand motion data, 2023.

[46] Arvind Narayanan and Vitaly Shmatikov. Robust De-anonymization of Large Sparse Datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125, Oakland, CA, USA, May 2008. IEEE. ISSN: 1081-6011.

[47] John William Nelson. A virtual property solution: How privacy law can protect the citizens of virtual worlds, 2010.

[48] James F. O'Brien, Robert E. Bodenheimer, Gabriel J. Brostow, and Jessica K. Hodgins. Automatic joint parameter estimation from magnetic motion capture data. In *Proceedings of Graphics Interface 2000*, pages 53–60, May 2000.

[49] UK's Information Commissioner's Office. Audits of data protection compliance by uk political parties. https://ico.org.uk/media/action-weve-taken/2618567/audits-of-data-protection-compliance-by-uk-political-parties-summary-report.pdf. Online; accessed 17 May 2022.

[50] Fiachra O'Brolcháin, Tim Jacquemard, David Monaghan, Noel O'Connor, Peter Novitzky, and Bert Gordijn. The convergence of virtual reality and social networks: Threats to privacy and autonomy.

[51] Ken Pfeuffer, Matthias J. Geiger, Sarah Prange, Lukas Mecke, Daniel Buschek, and Florian Alt. Behavioural Biometrics in VR: Identifying People from Body Motion and Relations in Virtual Reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–12, New York, NY, USA, May 2019. Association for Computing Machinery.

[52] Bridget Poetker. The Very Real History of Virtual Reality (+A Look Ahead).

[53] Frank E. Pollick, Jim W. Kay, Katrin Heim, and Rebecca Stringer. Gender recognition from point-light walkers. *Journal of Experimental Psychology: Human Perception and Performance*, 31:1247–1265, 2005. Place: US Publisher: American Psychological Association.

[54] Alan Dexter published. Oculus will sell you a Quest 2 headset that doesn't need Facebook for an extra $500. *PC Gamer*, April 2021.

[55] Michael L. Hicks published. Despite Quest 2 sales success, Meta lost $10.2 billion on VR/AR last year, February 2022.

[56] Aria Rezaei, Chaowei Xiao, Jie Gao, and Bo Li. Protecting sensitive attributes via generative adversarial networks, 12 2018.

[57] A.A. Rizzo, T. Bowerly, C. Shahabi, J.G. Buckwalter, D. Klimchuk, and R. Mitura. Diagnosing attention disorders in a virtual classroom. *Computer*, 37(6):87–89, 2004.

[58] Luc Rocher, Julien M. Hendrickx, and Yves-Alexandre de Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models.

[59] Luc Rocher, Julien M. Hendrickx, and Yves-Alexandre de Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models.

[60] Black Rock. The metaverse: Investing in the future now. https://www.blackrock.com/us/individual/insights/metaverse-investing-in-the-future. Online; accessed 17 May 2022.

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

[62] Christophe Olivier Schneble, Bernice Simone Elger, and David Shaw. The cambridge analytica affair and internet-mediated research. *EMBO reports*, 19(8):e46579, 2018.

[63] NATO Science and Technology Organization. Guidelines for mitigating cybersickness in virtual reality systems. https://www.sto.nato.int/publications/STO%20Technical%20Reports/STO-TR-HFM-MSG-323/\protect\T1\textdollar\protect\T1\textdollarTR-HFM-MSG-323-ALL.pdf.

[64] We Are Social. Digital 2022: Another year of bumper growth. https://wearesocial.com/uk/blog/2022/01/digital-2022-another-year-of-bumper-growth-2/. Online; accessed 17 May 2022.

[65] Morgan Stanley. Metaverse: more evolutionary than revolutionary. https://www.morganstanley.com/ideas/metaverse-investing. Online; accessed 17 May 2022.

[66] SteamDB. Most played VR Games Steam Charts.

[67] Labaton Sucharow. Record-breaking $650 million settlement of biometric privacy lawsuit reached by labaton sucharow, edelson, robbins geller and facebook. https://www.labaton.com/cases/550-million-settlement-principle-of-biometric-privacy-lawsuit-labaton-sucharow-facebook. Online; accessed 23 May 2022.

[68] Latanya Sweeney. Simple demographics often identify people uniquely, 2000.

[69] Latanya Sweeney, Akua Abu, and Julia Winn. Identifying Participants in the Personal Genome Project by Name. *SSRN Electronic Journal*, 2013.

[70] Philipp Sykownik, Divine Maloney, Guo Freeman, and Maic Masuch. Something personal from the metaverse: Goals, topics, and contextual factors of self-disclosure in commercial social vr. In *CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[71] Philipp Sykownik, Divine Maloney, Guo Freeman, and Maic Masuch. Something personal from the metaverse: Goals, topics, and contextual factors of self-disclosure in commercial social vr. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[72] Pier Paolo Tricomi, Federica Nenna, Luca Pajola, Mauro Conti, and Luciano Gamberini. You Can't Hide Behind Your Headset: User Profiling in Augmented and Virtual Reality, September 2022. arXiv:2209.10849 [cs].

[73] VRChat. Network specs and tips. https://docs.vrchat.com/docs/network-details.

[74] Ian Warren and Darren Palmer. Crime risks of three-dimensional virtual environments, 2010.

[75] P. Werner, S. Rabinowitz, E. Klinger, AD. Korczyn, and N. Josman. Use of the virtual action planning supermarket for the diagnosis of mild cognitive impairment: a preliminary study. *Dement. Geriatr. Cogn. Disord.*, 2009.

[76] Andrea Stevenson Won, Jeremy Bailenson, Jimmy Lee, and Jaron Lanier. Ho-muncular Flexibility in Virtual Reality. *Journal of Computer-Mediated Communication*, 20(3):241–259, 01 2015.

[77] David L. Woods, John M. Wyma, E. William Yund, Timothy J. Herron, and Bruce Reed. Age-related slowing of response selection and production in a visual choice reaction time task. *Frontiers in Human Neuroscience*, 9, 2015.

A EXPERIMENTAL DESIGN

This section describes the experiment design in detail. Our experiment consists of puzzles located in VR rooms that the participants visit. The puzzles are artifacts that facilitate collecting privacy-sensitive variables that might not otherwise be evident. The rooms are themed as a virtual office. Before initiating the game, we explained to the participants that they would find the password by solving a puzzle, thereby "escaping" the room. As developing a full-fledged game with voice recognition or virtual password pads is out of scope, the participants spoke the passwords aloud so that the researchers could press a key and "teleport" them to the next room. We include five "noisy" rooms, i.e., rooms that do not serve the purpose of facilitating the measurement of sensitive information but help to mask the rooms that do. Nonetheless, noisy rooms habituate the player to the game mechanics, e.g., looking around the room or immersing the player further in the game. If the player gets stuck in one room, we press a key to teleport the participant to the next room. We request the users to remove their glasses or contact lenses for puzzles 23 and 24, measuring eyesight. While influencing players in such a way is not possible in a real scenario, these puzzles could at least identify the players who do not have good eyesight, i.e., they do not wear glasses/contacts when playing.

Puzzle 1: The first room introduces the player to the dynamics of the game, containing only a door and a poster with the word "*hello*", which is the password. Upon instinctively reading the word aloud, the player is teleported to the next room.

Puzzle 2: The second room contains a poster with the password "*face*". The player spawns facing the opposite wall of the poster; thus, we accustom the player to turn and explore the virtual environment to find the password and reinforce finding and speaking the password aloud.

Puzzle 3: Similarly, a poster depicts a captcha with the word "*velvet*."

Puzzle 4: The room contains several tables with monitors, on whose screens are letters spelling "*church*" appropriately ordered from left to right.

Puzzle 5: This room tests for color blindness. Similarly to puzzle 4, monitors display letters on Ishihara color test plates. Without colorblindness, the player would read "*daisy*"; with colorblindness, the player would read "*as*" instead. Each of these passwords unlocks the room.

Puzzle 6: There is a button on a table; upon pressing it three times, the three balloons next to the opposite wall pop sequentially, revealing the password "*red*".

Puzzle 7: The puzzle tests the short-term memory of the participants (MoCA memory). A whiteboard displays seven rows arranged vertically, each with fill-in blanks. The first two rows contain the already filled-in words "*VR*" and "*hello*", respectively. The last five rows correspond to the previous passwords from puzzles 2 to 6. Connecting the highlighted letters sequentially from up to bottom, the participant reveals the password "*recluse*".

Puzzle 8: To measure wingspan, we depict on a wall four human stick figures with different poses. The participant must mimic the poses on the wall to uncover the four letters of the password "*cave*". One of the poses is a T-pose, which facilitates wingspan measurement.
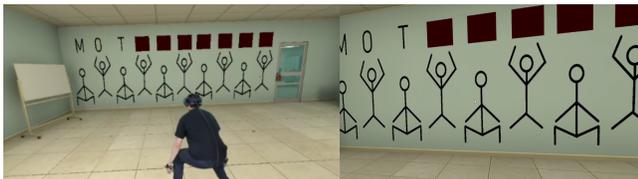
Puzzle 9: The participant must mimic the sequence of poses on the wall, a set of squats. For every squat, the participant uncovers two letters of the password "*motivation*". We correlate the distance traveled during the squats to fitness.

Puzzle 10: The (noisy) room depicts on a wall a pigpen cipher hiding the password "*deafening*".

Puzzle 11: The player presses a button on a table in time with a visual input, thereby revealing their reaction time.

Puzzle 12: The (noisy) room presents the password "*finally*" on the ceiling, habituating the user to look also upwards.

Puzzle 13: The room depicts the word "*apple*" in Hindi, Mandarin, French, Japanese, Russian, Spanish, Portuguese, and Arabic. The direction of gaze of the player when speaking the password reveals which language the participant recognizes.

Puzzle 14: This (noisy) room presents the sentence "*Everything you can do, I can do meta*" broken down vertically into five rows. To the left of each row, there is a shape. The last three shapes are the same (circles). To solve the puzzle, the participant must read aloud the words next to the first instance of the repeated shape "*I can.*"

Puzzle 15: Similarly to puzzle 14 and inspired by screen refresh rate tests [9], we present a number of balloons moving at different refresh rates. Depending on the refresh rate of the VR device, users cannot distinguish between some balloons.

Puzzle 16: To deploy the "naming" MoCA task, the room presents three whiteboards depicting three animals.

Puzzle 17: To measure an "attention" task from MoCA, we present a serial seven subtraction starting at 100, the password is the sequence of numbers that lead to the final answer: "65."

Puzzle 18: This room contains puzzle 7, thereby measuring delayed recall from the MoCA test.

Puzzle 19: This room pictographically recreates the MoCA abstraction test.

Puzzle 20 (no image): To complete this audio-only room, the participant must repeat aloud two recorded sentences after listening to them once, thereby measuring one of the language tests of the MoCA.

Puzzle 21: The (noisy) room depicts three pictures of a famous physicist—"*Albert Einstein*" is the password.

Puzzle 22: The room presents calendar days on a whiteboard with "*Today?*" as the header and without disclosing the year, month, weekday, or date, which prompts the participant to identify the date of the experiment, thereby measuring one variable of the orientation task in MoCA.

Puzzle 23: We measure whether a participant can read the text at a close distance. We write the sentence "*The code is equal to three times four*" in four lines on the screen of a monitor, each line becoming more diminutive than the above line.

Puzzle 24: Similarly, we measure whether a participant can read the sentence "*Life is better within the digital playground*" at a long distance.

B PARTICIPANT DISTRIBUTION

B.1 Demographics

Gender **50**
Male 26 (52.0%)
Female 24 (48.0%)

Age **50**
18–23 .... 24 (48.0%)
24–27 .... 20 (40.0%)
28–64 .... 6 (12.0%)

Nationality **50**
American 23 (46.0%)
Chinese 8 (16.0%)
Indian 6 (12.0%)
German 3 (6.0%)
Canadian 2 (4.0%)
Brazilian 1 (2.0%)
British 1 (2.0%)
Portuguese 1 (2.0%)
Spanish 1 (2.0%)
Swiss 1 (2.0%)
Undisclosed 3 (6.0%)

Ethnicity **50**
Asian 30 (60.0%)
White 14 (30.0%)
Black 3 (6.0%)
Hispanic 3 (6.0%)

Income **50**
≤ $25k 20 (40.0%)
$25k–$50k 15 (30.0%)
$50k–$100k 7 (14.0%)
≥ $100k 3 (6.0%)
Undisclosed 5 (10.0%)

Disability Status **50**
None 46 (92.0%)
Mental 3 (6.0%)
Physical 1 (2.0%)

Languages **50**
Chinese 20 (40.0%)
Spanish 14 (28.0%)
French 13 (26.0%)
Hindi 7 (14.0%)
None 6 (12.0%)
Portuguese 2 (4.0%)
Arabic 1 (2.0%)

B.2 Biometrics

Height **50**
150 cm – 165 cm 18 (36.0%)
166 cm – 175 cm 16 (32.0%)
176 cm – 189 cm 16 (32.0%)

Wingspan **50**
100 cm – 169 cm 21 (42.0%)
170 cm – 179 cm 18 (36.0%)
180 cm – 191 cm 11 (22.0%)

Longer Arm **50**
Left 26 (52.0%)
Right 18 (36.0%)
Same 6 (12.0%)

Reaction Time **50**
> 250 ms 27 (54.0%)
< 250 ms 23 (46.0%)

IPD **50**
< 63 mm 26 (52.0%)
63 mm – 66 mm 21 (41.0%)
> 66 mm 3 (6.0%)

Fitness **50**
Moderate 32 (64.0%)
High 10 (20.0%)
Low 8 (16.0%)

Colorblindness **50**
None 48 (96.0%)
Deuteranopia 2 (4.0%)

Hyperopia **50**
None 28 (56.0%)
Severe 13 (26.0%)
Mild 9 (18.0%)

Myopia **50**
Severe 32 (66.0%)
None 14 (28.0%)
Mild 4 (8.0%)

MoCA **50**
Pass (> 26) 43 (86.0%)
Fail (≤ 26) 7 (14.0%)

Handedness **50**
Right 47 (94.0%)
Left 3 (6.0%)

B.3 Environment

Location **50**
Location A 26 (52.0%)
Location B 20 (40.0%)
Location C 4 (8.0%)

Room Size **50**
$5 \ m^2$–$7 \ m^2$ 24 (48.0%)
$> 8 \ m^2$ 20 (40.0%)
$< 5 \ m^2$ 4 (8.0%)

Duration **50**
≤ 15 min 22 (44.0%)
16 min–20 min 18 (36.0%)
20 min–30 min 10 (20.0%)

Device **50**
Vive Pro 2 .... 26 (52.0%)
Oculus Quest 2 .... 21 (42.0%)
HTC Vive 3 (6.0%)

C   SOURCES OF GROUND TRUTH

Gender . . . . Self-Reported
Age . . . . .Self-Reported
Nationality . . . . .Self-Reported
Ethnicity . . . . .Self-Reported
Income . . . . Self-Reported
Disability Status . . . . Self-Reported
Languages . . . . Self-Reported
Height . . . . Stadiometer
Wingspan . . . . .Measuring Tape
Longer Arm . . . . .Measuring Tape
Reaction Time . . . . .Application
IPD . . . . Pupilometer
Fitness . . . . .Self-Reported
Colorblindness . . . . .Self-Reported
Hyperopia . . . . Self-Reported
Myopia . . . . .Self-Reported
MoCA . . . . .Administered
Handedness . . . . Self-Reported
Location . . . . GPS
Room Size . . . . Measuring Tape
Duration . . . . .Chronometer
Device . . . . .Observed