A Proposed Security Policy for the Metaverse

Shabnum (Sukhi) Gulati, Cooper Jones, Kelsey Merrill

May 10th, 2022

Abstract

This paper seeks to define a security policy for the metaverse. The paper begins by defining the metaverse, as the term has been used recently to refer to a wide spectrum of technologies. Then we present background information on known challenges faced by current metaverse implementations as well as existing legal mechanisms for protecting users online. Having explored the potential harms of a metaverse, we proceed to present a security policy for the metaverse. We analyze the policy by presenting potential legal and technical enforcement mechanisms. We conclude with a discussion of suggested future work. As the world rapidly adopts new paradigms of online interactions, the risk to consumers evolves in step. The metaverse provides many avenues for malicious actors to impact others on the platform. Platform providers and governmental entities must be willing to invest in new mechanisms for protecting metaverse users.

I. Metaverse Definition

While the concept of a “metaverse” has become popularly cited, there is little consensus on what exactly the metaverse is. One researcher crafted a paragraph-long definition which explains the metaverse as a “post-reality universe” consisting of an “interconnected web of social networked immersive environments in persistent multiuser platforms” [29]. News articles use the metaverse to refer to not just specific technologies, but often a mentality. Wired cites the Metaverse as not one specific technology, but rather a broad shift in how we interact with technology [33]. More blockchain focused lenses require that the metaverse has a fully functioning economy [9].

For the purposes of our security analysis, we are choosing a relatively limited version of the metaverse. We will focus on virtual-reality environments in which users can interact with other users in a richly interactive environment. Note that our definition focuses on metaverse implementations with virtual reality (VR) experiences. These platforms need not be exclusively VR. They can be accessed with a phone, but they should have a VR entry point. Another important component of our definition is the focus on metaverse implementations which enable social interactions in a computer-generated environment, as these social interactions present unique behavioral risks.

Our definition excludes popular implementations of block-chain metaverses such as the Sandbox coin due to the lack of VR access point. Nor do we necessitate commercial exchange. We will not perform a security analysis which covers block-chain based implementations of the metaverse or risks of metaverse implementations which include commercial exchange. We've drawn this line to start with a smaller scope for our security policy. Our definition of the metaverse includes, as a few examples, Meta's Horizon Worlds, VRChat, and Rec Room.

II. Background: Existing Metaverse Implementations & Challenges

In the past several years, we have seen an explosion in the number of metaverse implementations offered to consumers. While many of these applications cater to different demographics, they all suffer from very similar social problems. To understand potential societal harms of the metaverse, we analyzed behaviors on applications in-scope (such as Meta's Horizon Worlds), but we also looked at applications which might not fully fit our definition but still provide valuable insight (such as Second Life). These problems, ranging from harassment to child grooming and radical political indoctrination, serve as the focus for the rest of this section.

Harassment

Since the beginning of the internet, the combination of strong anonymity and weak enforcement of community policies has been a perfect storm for malicious actors to behave in any way they see fit with little chance of negative repercussions, often leading to a culture of harassment if left unchecked. Just like the real world, this harassment comes in many forms, but for the sake of this paper, we will focus predominately on two forms: sexual harassment and intentional destruction of property.

Sexual harassment is an issue that has long plagued Meta, and its adults-only Horizon Worlds app is no different. Since the early stages of its beta test, reports of inappropriate conduct such as groping have proliferated, with one user stating, "Not only was I groped last night, but there were other people there who supported this behavior which made me feel isolated in the Plaza" [10]. While Meta has some human moderators in Horizon Worlds, The Washington Post writes that there are never enough moderators to handle inappropriate events, saying: "moderators are sparse except in the app's few most populated spaces, and some users say they rarely intervene proactively or enforce the app's age restrictions" [30]. In the previously mentioned groping event, the victim states "I think what made it worse, was even after I reported, and eventually blocked the assaulter, the guide in the plaza did and said nothing. He moved himself far across the map as if to say, you're on your own now." Moderation is not a perfect solution as the moderators can continue to allow inappropriate behavior if there are even enough moderators to effectively intervene whenever necessary. Meta has recently begun attempting to curb the widespread harassment problem on their platform by adding a two-foot radius 'personal boundary' to each avatar which other avatars cannot enter [34].

Another form of harassment that legacy metaverse Second Life has had to address on multiple occasions is “Griefing”. Griefing is a form of virtual bullying resulting in a negative experience for the victim. While this attack can take several forms, the most common methods include “visually [vandalizing] the space by clouding it with large virtual objects that carry disparaging text or [graphics]” [22] and Denial of Service attacks that are caused by a large number of attackers flooding a virtual location or event, often with computationally intensive virtual objects, in order to overload the server and crash that virtual location. Griefing involves exploiting aspects of the virtual environment. In the vandalization example above, the malicious actor is taking advantage of their ability to spawn enough virtual objects to disrupt the space. Given the complexity of metaverse implementations, many different mechanisms and abilities may be vulnerable to Griefing.

Child Grooming

A problem social networks have faced since their invention is how to prevent children from interacting with potentially malicious adults. While there have been various attempts at age verification techniques, no platform has found an effective way to completely protect children from the dangers of the online world. Meta’s Horizon Worlds is no different. Even though it is marketed as an adults-only world and theoretically prevents children under the age of 18 from signing up, in practice, Meta has been unable to close all the loopholes available to kids who want to gain access to the platform. Several reports have surfaced concerning the large number of children using the app. One of these reports comes from a reporter exploring Horizon Worlds who was approached by a 9-year-old child who admitted that he was using his parents’ Oculus VR headset [30]. This event emphasizes the difficulty of locking down these online spaces for children. This is just one example of a widespread issue that currently has no good solution due to the inherent anonymity and freedom these platforms provide their users.

Radical Political Indoctrination

As far back as 2008, researchers have been concerned about the potential to use metaverse platforms as tools for political gain [28]. Focusing on Second Life, these researchers claimed that due to the similarity in-game interactions have to real world interactions, Second Life can provide extremist groups the opportunity to share their socio-political ideas with unsuspecting netizens, organize potential terrorist attacks, and provide ease of communication for better coordination. As was made clear during the events of January 6th at the United States’ Capital [41], when social media platforms ineffectively moderate the content on their platform, malicious agents can capitalize on the opportunity to foment extremism, amplify misinformation and coordinate potentially violent acts. Given the similarity of the metaverse to real life, the importance of effective content moderation on the platform is only made more important in order to protect all users and society in general.

Data Privacy

Data privacy concerns are already high with the web we are familiar with today, and the immense potential for data collection in the metaverse only heightens these concerns. The Wall Street Journal details some of the even more invasive data that could be collected in a virtual reality environment [38]. Examples include eye movements, gait, and tone of voice, all of which could provide insight into someone's underlying thoughts and emotions. This biological data is an addition to all the data we are more used to seeing generated, such as conversations with other users. Data privacy protections are especially important in the metaverse since users may not realize the extent of body language data they are giving away by simply existing in the space.

III. Background: Proposed & Actual Legislative Constraints

The following provides background on legislation relevant to a metaverse security policy. We are restricting the scope of this paper to a United States context, so everything discussed here is US legislation.

Section 230

Section 230 of the Communications Decency Act protects providers of "interactive computer services" from liability for content produced by third parties [23]. The law became effective in 1996, and it is often credited with aiding the development of the user-content rich internet we know today. "Interactive computer services" is interpreted very broadly to include almost any online service. Choosing to moderate content does not create liability for platforms under Section 230. Platforms can moderate content not at all, a little bit, or a lot, and they are not legally responsible for those moderation decisions or the content they decide to leave up.

The implications for the metaverse are similar to that of a traditional social media platform. Metaverse providers are not liable for any user-generated content on their platform, and they have no responsibility to moderate. They are also free to moderate in whatever manner they see fit.

As concerns with internet content grow, especially with the spread of sensationalist content or misinformation on social media, Section 230 protections have become highly controversial. The metaverse heightens these concerns. Some worry that the near-reality nature of the metaverse and an even greater movement of our lives online without moderation and regulation will create an environment full of offensive, illegal, unsafe, and untrue content [27]. Others worry that more aggressive regulation will hinder the metaverse's ability to develop [25]. The sheer amount of content created on the metaverse all the time creates an enormous moderation challenge that may be simply impossible for even large platforms, let alone smaller competitors.

The controversy around Section 230 has led to several proposals for its replacement. The following is a brief review of some proposed legislation:

- ➢ EARN IT Act (Senator Graham, 2022)
  The EARN IT Act removes Section 230 protection for CSAM (child sexual abuse material) [11]. If passed into law, providers could be held liable for the sharing or distribution of CSAM on their platform. The bill includes a carve-out for encrypted services.

- ➢ SAFE TECH Act (Senator Warner, 2021)
  The SAFE TECH Act removes Section 230 protection for material that the platform is paid to make available, the most common example being advertisements [35].

- ➢ JAMA Act (Rep Pallone, 2021)
  The Justice Against Malicious Algorithms Act removes Section 230 protections for large providers that use algorithms to make recommendations and "knowingly or recklessly makes a personalized recommendation that materially contributes to a physical or severe emotional injury to a person" [21].

- ➢ Protecting Americans From Dangerous Algorithms Act (Rep Malinowski, 2021)
  This act removes Section 230 protections for providers that use recommenders and recommend content that interferes with civil rights or promotes acts of international terrorism [31].

All of these examples remove some Section 230 protections and would require much more moderation on the part of metaverse providers and also additions to a metaverse security policy. They could also require changes to any recommender algorithms or to advertising infrastructure.

Children's Online Privacy Protection Act (COPPA)

The Children's Online Privacy Protection Act of 1998 (COPPA) provides online privacy protections for children under the age of thirteen. Under COPPA, online services covered by the law cannot collect information from children under the age of 13 without verifiable parental consent [6]. The law provides specific mechanisms that are considered verifiable parental consent, including calling a toll-free phone number or returning a signed form.

Not all online services are covered by COPPA. COPPA only applies to services "directed to children under 13," or services directed at a general audience but that the provider has "actual knowledge" that children under 13 are using their platform. The "actual knowledge" standard has been interpreted weakly. Many social media platforms get around COPPA protections by listing in their terms of service that users must be at least 13 years old and possibly asking users to enter their birthdate [14]. These terms of service clauses are often ignored, and children can easily lie about their birthdate to make an account, a known common practice. Yet, having the age clause and/or the birthdate checker is sufficient to avoid COPPA liability.

As mentioned above, children’s use of adult spaces in the metaverse is a major concern. As written, COPPA does not comprehensively address this concern. In the aforementioned example of children getting on to Meta’s 18+ Horizon Worlds, COPPA would not provide protection for those children since the platform is aimed at adults.

The failings of COPPA are well-known, and a replacement was proposed by Senator Markey in 2021. The Children and Teens’ Online Privacy Protection Act (CTOPPA) raises the age of consent for data collection from thirteen to sixteen, bans targeted advertising to children, and replaces the “actual knowledge” clause with “constructive knowledge” [5]. Constructive knowledge is defined as “directly or indirectly collects, uses, profiles, buys, sells, classifies or analyzes (using an algorithm or other form of data analytics) data” about the ages of users [15]. If this bill becomes law, the metaverse will have a much higher obligation to either keep children off their platforms or provide them the mandated privacy protections.

Data Privacy Legislation

Though there is no comprehensive data privacy legislation at the federal level in the US, California, Colorado, and Virginia have all passed state-level data privacy laws, and other states are expected to follow. California was the first US state to enact such legislation, and its laws have been used as a model for other states. The scope of Colorado and Virginia's data privacy laws are a bit more limited than California’s, so we’ll go into further detail only on California’s legislation to illustrate the content of these digital privacy protections.

California passed the California Consumer Privacy Act (CCPA) in 2018 after the threat of a ballot initiative, and it went into effect in January 2020. However, due to activist frustration with how regulators interpreted and implemented the CCPA, the California Privacy Rights Act (CPRA) was passed by ballot initiative in 2020. The CPRA will go into effect in January 2023.

The CPRA gives consumers the right to opt out of the sharing and sale of their personal data, correct inaccurate personal information, and receive transparent information about businesses data practices [4, 13, 17]. In order to be compliant with the CCPA/CPRA, the metaverse security policy will need to include statements about user data privacy and associated rights.

IV. Security Policy Proposal

We wrote this security policy with no specific metaverse platform in mind, rather it is meant to be general as to apply to any platform. Thus, any security policy for a specific implementation should expand upon this taking into account platform-specific design.

Agents

- Users - any person who utilizes the metaverse
    - Child users - users 15 and under
    - Adult users - users 16 and over
- Platform provider - the entity that develops, maintains, and runs the metaverse
- VR headset provider - the entity that distributes VR headsets used to access the metaverse
- Parents of child users - the legal guardian of child users

Policies

- All users
    - Authenticity
        - Account owners or users authorized by account owners should be the only users who can use that account
    - Access Control
        - All users should be able to read public user-generated content.
        - Users should only be able to read content in private channels they are participating in.
    - Consent & Awareness
        - Avatars should only be able to virtually touch other users with consent.
        - Users should be able to opt out of unwanted audio or visual sensations.
        - Users should clearly understand which behaviors are allowed and disallowed.
        - Users should maintain enough awareness of their physical surroundings to avoid injury to themselves and others.
    - Moderation
        - Users can "block" other users such that blocked users can't be seen, heard, or felt.
        - Users should be able to report content that violates communicated policies / community standards.
        - Users should not be able to engage in any form of sex trafficking or sexual harrassment.
        - Users should not be able to engage in any form of hate speech.
- Adult users
    - Access Control
        - Adults should not be able to access child-only spaces.
    - Consent

        - Users should have informed control over what personal data is collected and/or stored
        - Users should not be personally identifiable unless the platform is explicit that users will be personally identifiable by design."
- Child users
    - Access Control
        - Children should not be able to access adult-only spaces
- Platform provider
    - Access Control
        - Platforms should not be able to read content in explicitly private forums, regardless of the number of participants, without the user's knowledge.
        - Platforms should not, with the exception of unpredictable advancement in technology, be able to decipher content which they claim to the user to be encrypted.
    - Availability
        - The platform should be persistently available.
    - Moderation
        - Platforms should be able to remove public user-generated content per their communicated moderation policy.
        - The platform should be able to read public user-generated content.
        - Platforms should remove public content containing misinformation or hate speech that reaches a large number of people.
- VR headset provider
    - Access control
        - VR headset providers should only have access to data required for headset functionality.
- Parents/family of child users
    - Access Control & Moderation
        - Parents should have informed control over what personal data is collected and/or stored about their child.
        - Parents should be able to access and monitor their child's account.
        - Parents should know when and with whom their child shares personal information on the metaverse.

Definitions

Here we provide definitions, or caveats, for a few terms mentioned in the security policy which may be ambiguous:

Access Control: Regulating who has access to certain data or system resources.

Consent: For the purposes of our policy, “consent” refers to an explicit “yes” provided by the consentee in response to a question being asked by another agent. The consentee should be aware they are being asked for consent and what they are consenting or not consenting to. Consent must be able to be withdrawn at any time.

Awareness: Awareness is difficult to define. We won’t work through the nuances of what it means to be aware in this paper. However, whoever is asking for consent should be able to show they provided the user enough information to reasonably assume they were aware.

Moderation: Moderation refers to the practice of monitoring actions taken by and content generated by users on the platform in order to detect violations of community guidelines. The platform provider is expected to take some enforcement action when a violation is detected. Enforcement options are platform-dependent.

Community Guidelines: Community Guidelines refer to a collection of rules that an online application, usually one with a social or community component, sets for users of the application to follow.

Child-Only Spaces: Only users under age 16 are allowed in child-only spaces. Child-only spaces should be more strictly moderated than adult spaces as well as have parental control integration. Adult-only spaces, by contrast, do not have parental control access.

V. Security Policy Analysis: Policy Enforcement Mechanisms

This section will give an overview of legal and social mechanisms which can be used to enforce and encourage deployment of measures described in the security policy.

Community Guidelines

Platforms typically define community guidelines (defined above). Platforms may enforce community guidelines through mechanisms such as account suspensions or bans. Some of our security policies can be seen as suggested inclusions to community guidelines. For example:

- Users should not be able to engage in any form of sex trafficking or sexual harrassment.
- Users should not be able to engage in any form of hate speech.
- Platforms should remove public content containing misinformation or hate speech that reaches a large number of people.

By including such terms in community guidelines, platforms make it clear that they have a basis upon which to hinder or remove someone’s account. While our security policy does not intend to be an exhaustive list of all things that should be included in community guidelines, we’ve isolated a few harms (hate speech, sexual harassment, and misinformation) which we think should always be in community guidelines. We chose these two harms based on their unique potential for psychological damage in the metaverse, regardless of game mechanics. A notable exclusion, for example, is vandalism. Although vandalism of someone's metaversal space can be upsetting, it may be an essential part of a game. If a user expects their property to be vandalized and it is not linked to any property outside of the game, then the nature of vandalism is changed. Hate speech, sexual harrasment and widespread misinformation campaigns, we posit, should not be excused as part of any kind of game dynamic.

Comprehensive Federal Privacy Legislation & Data Handling Guidance

As many before us have argued, the US has a need for comprehensive privacy legislation. This will help enable the following specific aspects of our security policy:

- Platforms should not be able to read user conversations without the user's knowledge.
- Users should have informed control over what personal data is collected and/or stored.
- Users should not be personally identifiable unless the platform is explicit that users will be personally identifiable by design.
- VR headset providers should only have access to data required for headset functionality.

While companies can encode such measures in their own policies, there is little incentive to do so and consumers have little individual recourse to hold companies accountable. Drawing inspiration from abroad, the GDPR has strict requirements for the basis of processing personal data [20]. Such a foundation is particularly important with respect to any implementation of the 'metaverse'. While there has been much debate regarding the efficacy of the GDPR, due to the backlog of unresolved cases, the fines levied have produced observable changes in the care with which companies handle personal data [18].

Any such legislation should include clauses sensitive to the forms of data collection enabled by virtual reality. Specifically, eye tracking and gait analysis. We'll call this data "Body Data." More so than text-based user generated content, users are not as likely to be aware of the extent to which their Body Data is being revealed to others in the metaverse. As a result, information the user intends to keep private such as moods or disability status could be inadvertently revealed. Body Data is unique, as it is not sufficient to only legislate how the data processor collects the data since the data can be easily collected by an adversary by observing a realistically rendered avatar. For example, even if Meta stores Body Data anonymized through differential privacy, an adversary could collect this data tied to an original account if Meta chooses to render avatars as realistically as possible. To prevent an adversary harvesting this data, it could be obfuscated with random noise not just at collection time but at render time as well. Alternatively, users could opt out of realistic avatar-rendering all together. Legislation, then, should mandate that any externally personally identifiable data should be sufficiently anonymized not just in collection by the service provider but at render time. The legislation need not be prescriptive of the exact technical method of anonymization and may provide exemptions with explicit user consent.

Any digital privacy legislation should also guard against risks introduced by hardware providers. Abuse of data access by headset providers is a growing concern. In fact, it was revealed that Samsung Smart TVs analyzed users' viewing habits to serve them better ads [1]. It is easy to imagine a scenario where VR headset providers analyze what a user is doing in the metaverse in order to profit off of users without their consent. In order to protect user's privacy and prevent data they generate from being used unbeknownst to them, we propose that the only data VR headset providers should be able to access is that which is strictly necessary to the functionality of the headset. We believe that any analysis of user behavior beyond what is required for functionality is an aggressive breach of user privacy and as such we seek to ban this practice outright, putting users' privacy rights before the profit interests of the companies providing the VR headsets. Restrictions as to what data hardware providers can collect should be included in comprehensive data privacy legislation. Given the difficulty of restricting hardware access to data technically, we feel a legal deterrent would be the most effective solution to this data privacy harm.

CTOPPA Endorsement

We strongly support Senator Markey's amendment to COPPA. Recall, the proposed addendum to COPPA raises the age of consent for data collection from thirteen to sixteen, bans targeted advertising to children, and replaces the "actual knowledge" clause with "constructive knowledge." Two of the three proposed changes are particularly relevant to our understanding of harms in the metaverse. First, raising the age from 13 to 16 protects more children. Child grooming, a key potential harm of the metaverse, is a risk which does not stop at age 13 [7]. Second, moving to a standard of constructive knowledge helps mitigate the incentive which corporations have to turn a blind-eye to children on the platform. Age verification measures, some of which we'll touch upon in the technical section, can be costly to implement and dissuade participants. A legislative incentive is likely to be more effective than market incentives to drive implementation as a result.

Section 230 Amendments

We do not currently endorse any of the proposed amendments to Section 230, as we tend to believe that the limited liability afforded to internet intermediaries provided by Section 230 has generally allowed internet application development to positively grow and innovate. However, holding platforms legally accountable for speech on their platforms instead of relying on community guideline enforcement would certainly be a stronger legal mechanism. As we have in our security policy, we would encourage regulators to consider scale. A Section 230 amendment which holds platforms accountable for misinformation after it has reached a certain number of people or a certain percentage of the platform, for example, may be both feasible to enforce and prevent societal harm. By contrast, an amendment declaring platforms liable for any form of misinformation is likely to lead to ineffectual enforcement due to the sheer scale of content.

VI. Security Policy Analysis: Technical Enforcement Mechanisms

Technical Security Mechanisms

In this section, we present technical security mechanisms for each of the statements in our security policy.

Account Verification

- Account owners or users authorized by account owners should be the only users who can use that account

In order to verify that the person that set up and owns a given account is actually the one using that account, we propose biometric scanning through the VR headset. We propose a form of biometric scanning implemented with the following privacy guardrails. First, biometric scanning should simply be to check for a match with the image configured at set-up. No other personally identifiable information should be stored alongside the biometric data. Second, the biometric scanning should be completely client-side. No biometric data should be transmitted to the service provider or over the network at any time. The user's device should store the image configured on start-up and simply match new scans to that image on device. The second point about storing the data only on the device is especially important. Since biometric data is often unique to an individual, there are still concerns with it being leaked even if it's not stored alongside other personally identifiable information. With both of these guardrails in place, we intend to mitigate the privacy concerns often associated with biometric scanning.

One possible implementation suited for VR is iris scanning. Iris scanning uses infrared light to scan the eye and identify patterns unique to every individual, similar to a fingerprint [26]. Since the VR headset is worn over the eyes, iris scanning fits naturally into this context. Samsung successfully implemented iris scanning for security in its S8 and S9 series smartphones [19], and a 2021 study found iris scanning to be accurate enough to identify participants in a medical clinical trial [39], so we feel confident it could be implemented in a VR setting.

Age Verification

- Adults should not be able to access child-only spaces.
- Children should not be able to access adult-only spaces.

In order to make sure that users are only accessing age-appropriate spaces, the platform can associate the age of the user with the account. In order to ensure correctness, the platform should conduct some age verification. Some commonly used techniques are listed below [2]:

- Have users submit a photo of their ID and use optical character recognition (OCR) to read the birth date on the ID. Sometimes, facial matching with the photo on the ID is also used to verify that the submitter actually matches the ID.
- Users enter their name, address, and date of birth, and the platform runs a check with a credit agency to determine that users' age.
- Users enter a valid credit card. Most of the time, banks will not issue credit cards to minors, so this provides strong evidence that someone is over 18.

This list has a number of issues. First, these techniques are really only effective for verifying the age of people 18 and over. Users under 18 are much less likely to have a photo ID displaying their birth date or have a credit score with a credit agency, and the credit card check cannot distinguish between a 12-year-old and a 13-year-old because neither will have credit cards. Also, these techniques come with privacy concerns as they require the user to identify their real-world self and give up personally identifying information to the platform. Checking an ID could be done completely client-side, thereby assuaging many privacy concerns, but this still does not solve the problem for users without a photo ID.

Another option is to use machine learning to determine the age of the user based on their facial features. This approach has privacy advantages over the above list because the user doesn't have to reveal any more information than they would otherwise - simply wear the VR headset. Also, this approach would work for users of any age and doesn't rely on social adulthood indicators. There is active research in this area, but the technique still has a ways to go until it is accurate enough to be used in this way. Recent studies put the accuracy of this approach at around 70% [12, 24, 32]. Furthermore, bias in machine learning and particularly facial recognition is a documented problem, so this accuracy might be much lower for some users [3]. Though a promising option for age verification in the future, this technique is not ready for deployment today.

Short term, we would propose credit-card based age verification when possible. While not perfect, users often have to provide payment information to VR app ecosystems regardless. Credit-card bearers, who will generally be adults, would then be needed to assist children at initial account set-up and thus prevent children from lying about their age.

Availability

- The platform should be persistently available

Because of the unique way the metaverse functions as a sort of alternate life for many of its users, we felt enforcing that the platform will always be available with minimal disruptions was important to protecting users as they build up their new "life" in the metaverse. It's hard to anticipate all possibilities for how a system might become unavailable, so we will touch specifically on attacks which came up in our research on past harms. Specifically, we read about Griefing attacks where users would leverage loopholes in the system to architect Denial of Service attacks [22]. For example, users may be able to have their avatar carry visible objects. Malicious users would pick an object with large compute time to render then enter a virtual, public space in order to degrade performance for other users who would experience computation lags as their devices attempted to render the adversary's object. Of course, metaverse providers are also susceptible to external DDoS attacks. Such attacks can be prevented using auto scaling functionality and DDos protections from cloud providers as well as computation caps on the feature being exploited [42]. This will prevent an area with a large amount of users from experiencing degraded performance if the computational toll begins to exceed what the normal server setup can handle while also ensuring that a malicious user flooding the system with network packets is unable to drown out legitimate network traffic from users attempting to use the system. In the case that the server auto scaling is not responsive enough or adequate for the computational resources required to render a scene, we propose implementing logic to prevent unnecessary visuals from rendering while still preserving essential gameplay mechanics such as background rendering, movement, and communication with other users in order to take the load off of the servers. However, we've only covered a few types of attacks here. It will be essential for metaverse platforms to take stock of how adversaries can exploit platform features in order to degrade system performance.

Avatar Separation

- Avatars should only be able to virtually touch other users with consent
- Users should be able to opt out of unwanted audio or visual sensations.
- Users should not be able to violate sex trafficking and sexual harrassment law on the platform

In order to prevent unwanted "physical" interactions in the metaverse, the platform can implement a "personal space bubble." Another avatar cannot enter someone else's personal space bubble without explicit consent from the user in question. For example, if one avatar wanted to give another avatar a hug, they could ask to enter that person's personal space, and then that person would have the opportunity to either allow the interaction to continue or deny the request. Meta has already implemented this for their Horizon Worlds platform following reports of virtual groping [34].

To prevent unwanted audio or visual sensations, users should be given the option to "mute" or "hide" any other user or content being displayed by that user. This can be seen as an audiovisual form of avatar separation, but with a block-list model instead of the allow-list model we've proposed for physical touch. Since audiovisual sensations are core to constructing the metaverse environment, an allow-list model would not be feasible even though it would be more secure against harm.

Parental Controls

- Parents should have informed control over what personal data is collected and/or stored about their child
- Parents should be able to access and monitor their child's account

In order to give parents adequate supervision and control over their child's activity in the metaverse, we propose a parent "joint account." The joint account would be connected to the child's account and include relevant statistics, notices, and parental controls specific to the platform. It would also allow the parent to configure data privacy settings for their child as well as enable parental moderation of content consumed and generated by their child. Such parental controls should be technically feasible since platforms have access to all the data. We foresee two potential challenges. First, encrypted spaces cannot be moderated. We touch on this later in private conversations where we suggest a weaker form of encryption (client-server as opposed to end-to-end) for spaces including children. The second challenge is that it may be difficult to create an effective moderation UX if the child is generating and consuming large amounts of content. If the encryption is configured as described in our first point, the platform can run content analysis to screen for particularly concerning content and flag this to parents. Overall, the type of parental control software depends a lot on the features offered by the platform, but we believe it to be technically feasible.

In order to correctly identify a child's parent, we rely on the FTC's published guidelines for getting parent's verifiable consent under COPPA [6]. Note that minors aged 16 and up are not defined as children under our security policy. A joint account would be mandatory only for users 15 or under. Although 16 and 17 year olds are at risk of child grooming, they also have a right to privacy. When does a child's right to privacy come before shielding them from risk? The answer is subjective and our choice of 16 is, admittedly, a bit arbitrary. We chose based on intuition and a survey of legislation. There is an important balance to strike when it comes to protecting children and still preserving their right to privacy.

Personal Data Collection / Human-Computer Interaction

- Users should have informed control over what personal data is collected and/or stored
- Users should clearly understand which behaviors are allowed and disallowed.
- Users should not be personally identifiable unless the platform is explicit that users will be personally identifiable by design.

In any online space, the primary user goal is not to avoid data collection but instead to complete the task at hand. This makes communicating privacy policies or community guidelines difficult from a human-computer interaction (HCI) perspective. There are also open questions around how to communicate policies and choices about data use or other complex or abstract topics to a user in a way that makes sense and allows them to actually conceptualize what is going on. These same questions will be important when considering how to truly give users "informed control" over data collection and make sure users "clearly understand" allowed and disallowed behaviors.

Like the rest of this research community, we have no clear and comprehensive solution right now. However, here are some best practices paraphrased from Schaub, Balekabo, Durity, and Cranor [36] that platforms should follow when communicating this kind of information to users and giving them choices about their data:

1. It is not acceptable to present users with a long "privacy policy" or "terms of use" during setup and expect users to click "I Agree." Users will not read it because it interrupts their pursuit of the primary goal, and it doesn't give users any sort of granular control.
2. Instead of communicating everything that a regulator might need to know to a user, use different notices for different contexts. Carefully choose what information users will need, and prioritize "surprises" in notices, or things a user would not expect.
3. Use layered and contextualized notices. Provide notice and ask for consent about specific things when it is relevant.
4. Design notices specific to the space, and do user testing. This will make sure the notices feel natural in context and are easy for users to interact with and understand.

Physical Awareness

- Users should maintain enough awareness of their physical surroundings to avoid injury to themselves and others.

In order to account for users' physical safety and the safety of those around them, users must be aware when they are about to come into contact with objects in the real world. The platform can assist in this by allowing users to set up their "play area" and then alerting users if they move outside of this area. Meta has implemented this feature in their Oculus VR headsets [37]. Generally, this is best enforced by the VR headset provider to ensure a consistent user experience across VR applications, since the need for physical awareness exists across VR applications.

Private Communications

- Users should only be able to read content in private channels they are participating in.
- Platforms should not be able to read content in explicitly private forums, regardless of the number of participants, without the user's knowledge.
- Platforms should not, with the exception of unpredictable advancement in technology, be able to decipher content which they claim to the user to be encrypted.
- Parents should know when and with whom their child shares personal information on the metaverse

In order to ensure confidentiality for private conversations between adult users, the platform can use end-to-end encryption. This way, it is computationally impossible for anyone other than the user “endpoints” to read the conversation, including the platform. A model implementation for end-to-end encryption is the Signal protocol, which is open source and has been widely deployed in the Signal messaging app and WhatsApp. It has also been extensively analyzed by academics and other researchers and shown to be secure [8].

For direct message conversations involving at least one child user, we prioritized child safety over privacy from the platform. Instead of end-to-end encryption for these conversations, we propose client-server encryption. This way, only the users engaged in the conversation and the platform can read messages. We then propose some automated scanning on the server to look for personally identifiable information like phone numbers, email addresses, or addresses, so parents can be alerted if their child shares this information with someone else in the metaverse. The technology for detecting personally identifiable information (PII) is readily available and commonly used in industry for applications such as removing PII from cloud data stores [40]. While the technology is currently only being applied to text, given that audio to text products are extremely mature, it is very feasible to hook these two technologies together in order to detect children sharing PII either over text or speech so that the platform can notify the parent. We would also endorse client-side scanning of image content sent in chats including children to flag potential CSAM in parental control software for parents to take appropriate action upon.

Tiered Moderation

- Platforms should be able to remove public user-generated content per their communicated moderation policy.
- Platforms should remove public content containing misinformation or hate speech that reaches a large number of people.

In an ideal world, content that violates community guidelines and/or contains misinformation or hate speech would not exist on the metaverse at all. However, requiring such fine-grained moderation is infeasible if the norm of low barriers-to-entry user-generated content is to be maintained. Platforms simply do not have the resources to moderate all public-facing content that users create.

Instead, we propose a tiered moderation system. As content reaches more people, it has more ability to do harm, and thus deserves more resources and attention from the platform. Tiered moderation means that content receives moderation scrutiny proportional to its reach. Tiered moderation systems enable users to continue posting and interacting with others online freely, but they prevent harmful content like misinformation and hate speech from reaching large audiences and doing widespread damage.

Twitter implemented a system like this leading up to the November 2020 elections, applying special moderation practices to accounts with more than 100,000 followers and posts with more than 25,000 likes [16]. We envision a system like this could function on the metaverse. For example, a user speaking to a large audience would also be listened to by a human moderator who could stop their speech if it contained content in violation of these policies, but a conversation between a few friends would not receive these resources. We will note that it's difficult to evaluate the effectiveness of these systems, but believe scaled moderation would present an improvement over the sparse moderation in existing metaverse environments.

No Novel Mechanisms Needed

- Users can "block" other users such that blocked users can't be seen or heard.
- The platform should be able to read public user-generated content.
- All users should be able to read public user-generated content.
- Users should be able to report content that violates communicated policies / community standards.

These policies have well established technical enforcement mechanisms or are trivially enforced. Paradigms exist to block users or report content. By its public nature, public user-generated content is visible to the platform and all users.

Covered by Policy Mechanisms

- Users should not be able to engage in any form of sex trafficking or sexual harrassment.
- Users should not be able to engage in any form of hate speech.
- VR headset providers should only have access to data required for headset functionality.

These policies are covered by legal mechanisms, and we are not proposing additional technical enforcement mechanisms.

VII. Conclusion

While we believe our proposed security policy makes progress in identifying user harms as metaverse implementations rise in popularity, it is clear from our proposed enforcement mechanisms that technical and social progress is required to effectively ensure the safety of metaverse users.

Technical areas of further development include: adopting an industry standard method for age and account verification, developing the technology to prevent users from having to give up sensitive personal information while still being able to have unique, personal experiences on the platform, increasing efforts to improve content moderation in real time, and formalizing VR-specific design patterns necessary to ensure the protection of all users such as respecting avatar's physical boundaries and separating users who should not normally interact in the real world. It is worth noting that, with the exception of end-to-end encryption, most of our proposed technical mechanisms are system-level techniques, not provable cryptographic techniques. They make it harder for an adversary to break the rules, but not impossible. They do not provide perfect enforcement of our security policy. However, we believe these improvements would go a long way in protecting users.

Our suggested areas of legal and social enforcement would require policymakers to take clear and decisive action in increasing protections for children online, provide clear guidance on platform-specific community guidelines, and pass comprehensive digital privacy legislation. The political viability of passing such legislation is rapidly changing. As demonstrated by our analysis of current proposals, there seems to be momentum around creating more effective technical legislation. We also see global progress, particularly in Europe, with the introduction of digital privacy legislation. We hope to see this momentum capitalized upon in the United States.

To accomplish any of the requisite technical or social progress, there is also a need to drive consensus on underlying values. Implicit in our security policy and proposed enforcement mechanisms are many judgment based privacy vs. security trade-offs. Perhaps most notably, defining users 16 and up as adults creates a line in the sand between when an individual's privacy should come before their legal status as a minor. Similarly, there are inherent risks with using biometric verification for VR headset users. While we propose technical mitigations for some of these risks (ie: client-side storage only), the willingness to incorporate biometric verification at all stems from an inherent value judgment that protecting children against potential harms is worth trusting a platform to handle this sensitive data correctly. There also exists some tension between the enforcement of sex trafficking law, sexual harrassment law, and other real-world laws and user privacy and the use of encryption. Our security policy has left some of this trade-off up to the platform. We do not require private conversations to be encrypted, but rather that any platform which claims to be encrypted is effectively encrypted. However, we also include a statement that platforms should not be able to read private user conversations without user knowledge which directly hinders a platform's ability to proactively investigate illicit behavior. Therefore, our policy still presents a preference for preserving user privacy over law enforcement needs. The debate regarding the appropriate amount of law enforcement access to online forums is far from settled.

We hope our policy proposal can serve as a baseline which platform providers, end users, and policymakers can build upon in order to implement widespread legal and technical protection mechanisms and ensure the safety of each and every user who desires to join this new era of interacting online. We also hope the proposal can shed light on how much more work is needed to enforce an effective security policy, especially as we've written our policy for a limited definition of the metaverse. While metaverse implementations gain popularity, user protections do not necessarily grow at the same rate without a conscious effort. Further interdisciplinary mobilization and analysis will help ensure that society can safely benefit from new frontiers in virtual reality experiences.

References

1. Anderson, T. (2020, September 30). *Who watches the watchers? Samsung does so it can fling ads at owners of its smart tvs*. The Register. Retrieved May 9, 2022, from https://www.theregister.com/2020/09/30/samsung_smart_tv_ads/
2. AVPA. (n.d.). *Age verification methods*. AVPA. Retrieved May 9, 2022, from https://avpassociation.com/avmethods/
3. Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Conference on Fairness, Accountability, and Transparency*.
4. *California Privacy Rights Act (CPRA)*. Perkins Coie. (n.d.). Retrieved May 9, 2022, from https://www.perkinscoie.com/en/practices/security-privacy-law/california-privacy-rights-act-cpra.html
5. Children and Teens' Online Privacy Protection Act (2021). bill.
6. *Children's Online Privacy Protection Rule: A Six-step compliance plan for your business*. Federal Trade Commission. (2020, July 17). Retrieved May 9, 2022, from https://www.ftc.gov/business-guidance/resources/childrens-online-privacy-protection-rule-six-step-compliance-plan-your-business
7. Choo, K.-K. R. (2009). (rep.). *Online child grooming: a literature review on the misuse of social networking sites for grooming children for sexual offences*.
8. Cohn-Gordon, K., Cremers, C., Dowling, B., Garratt, L., & Stebila, D. (2017). A formal security analysis of the signal messaging protocol. *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*. https://doi.org/10.1109/eurosp.2017.27
9. Cointelegraph. (2022, March 28). *What is metaverse in Blockchain? A beginner's guide on an internet-enabled virtual world*. Cointelegraph. Retrieved May 9, 2022, from https://cointelegraph.com/metaverse-for-beginners/what-is-metaverse-in-blockchain-a-beginners-guide-on-an-internet-enabled-virtual-world
10. Duffield, W. (2022, February 9). *A grope in Meta's space*. Techdirt. Retrieved May 9, 2022, from https://www.techdirt.com/2021/12/28/grope-metas-space/
11. Eliminating Abusive and Rampant Neglect of Interactive Technologies Act of 2022 (2022). bill.
12. Erbilek, M., Fairhurst, M., & Abreu, M. (2017). Age predictive biometrics: Predicting age from Iris characteristics. *Iris and Periocular Biometric Recognition*, 213–234. https://doi.org/10.1049/pbse005e_ch10
13. Fennessy, C. (2022, March 4). CPRA's top-10 impactful provisions. Retrieved May 9, 2022, from https://iapp.org/news/a/cpra-top-10-impactful-provisions/
14. Finnegan, S. (2020, January 9). How Facebook Beat the Children's Online Privacy Protection Act: A Look into the Continued Ineffectiveness of COPPA and How to Hold Social Media Sites Accountable in the Future. Seton Hall University.
15. Fitzpatrick, A. (2021, October 11). *"Children and teens' online privacy protection act" offers potential changes to Coppa Requirements*. Davis+Gilbert LLP. Retrieved May 9, 2022, from https://www.dglaw.com/children-and-teens-online-privacy-protection-act-offers-potential-changes-to-coppa-requirements/
16. Gadde, V. (2020, October 9). *Additional steps we're taking ahead of the 2020 US election*. Twitter. Retrieved May 9, 2022, from https://blog.twitter.com/en_us/topics/company/2020/2020-election-changes
17. George, D. (2021, January 28). *California's Consumer Privacy Rights Act: OPT-out rights and data profiling*. The National Law Review. Retrieved May 9, 2022, from https://www.natlawreview.com/article/california-s-consumer-privacy-rights-act-opt-out-rights-and-data-profiling
18. Heine, I. (2022, May 2). *3 years later: An analysis of GDPR enforcement*. Center for Strategic and International Studies. Retrieved May 9, 2022, from https://www.csis.org/blogs/strategic-technologies-blog/3-years-later-analysis-gdpr-enforcement

19. *How does the IRIS scanner work on Galaxy S9, Galaxy S9+, and galaxy note9?* The Official Samsung Galaxy Site. (n.d.). Retrieved May 9, 2022, from https://www.samsung.com/global/galaxy/what-is/iris-scanning/
20. Intersoft Consulting. (2019, September 2). *General Data Protection Regulation*. Retrieved May 9, 2022, from https://gdpr-info.eu/
21. Justice Against Malicious Algorithms Act of 2021 (2021). bill.
22. Lee, C. Y., & Warren, M. (2007). *Security issues within virtual worlds such as second life*. Research Online. Retrieved May 9, 2022, from https://ro.ecu.edu.au/ism/44/
23. Legal Information Institute. (n.d.). *47 U.S. Code § 230 - protection for private blocking and screening of offensive material*. Legal Information Institute. Retrieved May 9, 2022, from https://www.law.cornell.edu/uscode/text/47/230
24. Levi, G., & Hassncer, T. (2015). Age and gender classification using Convolutional Neural Networks. *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. https://doi.org/10.1109/cvprw.2015.7301352
25. Londoño, J. (2022, March 17). *The erosion of intermediary liability protections can end the metaverse before it even starts*. The Erosion of Intermediary Liability Protections Can End the Metaverse Before It Even Starts. Retrieved May 9, 2022, from https://itif.org/publications/2022/03/17/erosion-intermediary-liability-protections-can-end-metaverse-it-even-starts
26. Lynch, J., Schwartz, A., Gullo, K., Hussain, S., & Sheard, N. (2019, October 26). *Iris recognition*. Electronic Frontier Foundation. Retrieved May 9, 2022, from https://www.eff.org/pages/iris-recognition
27. MacCarthy, M. (2022, February 9). *Coming soon to a podcast, an App Store and a metaverse near you.... content moderation rules*. Forbes. Retrieved May 9, 2022, from https://www.forbes.com/sites/washingtonbytes/2022/02/03/coming-soon-to-a-podcast-an-app-store-and-a-metaverse-near-you-content-moderation-rules/?sh=33ca9bd57487
28. Mandal, S., & Lim, E.-P. (2008). Second life: Limits of creativity or cyber threat? *2008 IEEE Conference on Technologies for Homeland Security*. https://doi.org/10.1109/ths.2008.4534503
29. Mystakidis, S. (2022, February 10). *Metaverse*. MDPI. Retrieved May 9, 2022, from https://www.mdpi.com/2673-8392/2/1/31
30. Oremus, W. (2022, February 7). *Kids are flocking to Facebook's 'metaverse.' experts worry predators will follow.* The Washington Post. Retrieved May 9, 2022, from https://www.washingtonpost.com/technology/2022/02/07/facebook-metaverse-horizon-worlds-kids-safety/
31. Protecting Americans from Dangerous Algorithms Act (2021). bill.
32. Rahman Hassan, K., & Hadi Ali, I. (2020). Age and gender classification using multiple Convolutional Neural Network. *IOP Conference Series: Materials Science and Engineering*, *928*(3), 032039. https://doi.org/10.1088/1757-899x/928/3/032039
33. Ravenscraft, E. (2021, November 25). *What is the metaverse, exactly?* Wired. Retrieved May 9, 2022, from https://www.wired.com/story/what-is-the-metaverse/
34. Robertson, A. (2022, February 4). *Meta is adding a 'personal boundary' to VR avatars to stop harassment*. The Verge. Retrieved May 9, 2022, from https://www.theverge.com/2022/2/4/22917722/meta-horizon-worlds-venues-metaverse-harassment-groping-personal-boundary-feature
35. Safeguarding Against Fraud, Exploitation, Threats, Extremism, and Consumer Harms Act (2021). bill.
36. Schaub, F., Balebako, R., Durity, A. L., & Cranor, L. F. (2015). A design space for effective privacy notices. *The Cambridge Handbook of Consumer Privacy*, 365–393. https://doi.org/10.1017/9781316831960.021
37. *Setting up your play area and guardian*. Social Metaverse Company. (n.d.). Retrieved May 9, 2022, from https://store.facebook.com/help/quest/articles/in-vr-experiences/oculus-features/oculus-guardian/
38. Uberti, D. (2022, January 4). *Come the metaverse, can privacy exist?* The Wall Street Journal. Retrieved May 9, 2022, from https://www.wsj.com/articles/come-the-metaverse-can-privacy-exist-11641292206

39. Zola Matuvanga, T., Johnson, G., Larivière, Y., Esanga Longomo, E., Matangila, J., Maketa, V., Lapika, B., Mitashi, P., Mc Kenna, P., De Bie, J., Van Geertruyden, J.-P., Van Damme, P., & Muhindo Mavoko, H. (2021). Use of Iris scanning for biometric recognition of healthy adults participating in an ebola vaccine trial in the Democratic Republic of the Congo: Mixed methods study. *Journal of Medical Internet Research*, *23*(8). https://doi.org/10.2196/28573
40. Hegner, T. (2020, December 14). *How to identify PII in text fields and REDACT IT*. phData. Retrieved May 10, 2022, from https://www.phdata.io/blog/how-to-identify-pii-in-text-fields-and-redact-it/
41. Timberg, C., Dwoskin, E., & Albergotti, R. (2021, October 22). *Inside facebook, Jan. 6 violence fueled anger, regret over missed warning signs*. The Washington Post. Retrieved May 10, 2022, from https://www.washingtonpost.com/technology/2021/10/22/jan-6-capitol-riot-facebook/
42. Anderson, P. (n.d.). *Shield*. Amazon. Retrieved May 10, 2022, from https://aws.amazon.com/shield/ddos-attack-protection/