Harmful Design in the Metaverse and How to Mitigate it: A Case Study of User-Generated Virtual Worlds on Roblox

Yubo Kou
College of Information Sciences and Technology,
Pennsylvania State University, USA
yubokou@psu.edu

Xinning Gui
College of Information Sciences and Technology,
Pennsylvania State University, USA
xinninggui@psu.edu

ABSTRACT

Metaverse platforms such as Roblox have become increasingly popular and profitable through a business model that relies on their end users to create and interact with user-generated virtual worlds (UGVWs). However, UGVWs are difficult to moderate, because game design is inherently more complex than static content such as text and images; and Roblox, a game platform targeted primarily at child players, is notorious for harmful user-generated game such as Nazi roleplay games and gambling-like mechanisms. To develop a better understanding of how harmful design is embedded in UGVWs, we conducted an empirical study to understand Roblox users' experiences with harmful design. We identified several primary ways in which user-generated game designs can be harmful, ranging from directly injecting inappropriate content into the virtual environment of UGVWs to embedding problematic incentive mechanisms into the UGVWs. We further discuss opportunities and challenges for mitigating harmful designs.

CCS CONCEPTS

• **Human-centered computing** → Human computer interaction (HCI); Empirical studies in HCI; Interaction design; Empirical studies in interaction design.

KEYWORDS

Harmful design, user-generated virtual world, metaverse, design moderation

ACM Reference Format:
Yubo Kou and Xinning Gui. 2023. Harmful Design in the Metaverse and How to Mitigate it: A Case Study of User-Generated Virtual Worlds on Roblox. In *Designing Interactive Systems Conference (DIS '23), July 10–14, 2023, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3563657.3595960

1 INTRODUCTION

Persistent, three-dimensional (3D) virtual worlds (more often referred to today as the 'metaverse') have been popular in the past few decades. While traditional virtual worlds such as Second Life, often considered as the first metaverse [101], and World of Warcraft were created and moderated by professional developers in a centralized manner, the modern-day ambitions of tech companies such as Meta [71] and Microsoft [62] are putting more emphasis on facilitating and monetizing user-generated content (UGC) in their respective metaverse platforms. Pushing this idea even further, metaverse platforms such as Roblox have developed a business model that relies entirely on user-generated virtual worlds (UGVWs) by monetizing UGVWs and sharing revenues with UGVW creators [78]. Metaverse platforms like Roblox usually provide a whole development environment with low-barrier scripting languages such as Lua, making it easy for many end users, including children, to design and create UGVWs [75]. Roblox end users have the ability to interact with UGVWs as players and to create their own UGVWs as creators.

However, end users' empowered capabilities to design UGVWs on metaverse platforms have not come with adequate oversight. Roblox has regularly made headlines for harmful designs generated by its end users. In this paper, harmful design refers broadly to design patterns in UGVWs that can incur harm to players who interact with such UGVWs. For example, Roblox users have designed Nazi role-playing UGVWs [27], "condo games," which are UGVWs that could expose child players to sex acts and crude language [108], or UGVWs that recreate Nazi death camps [77] and mass shootings [14]. However, despite growing societal concern (e.g., [14, 55, 58, 80, 108]), the research community has paid limited attention to the issue of harmful design on metaverse platforms.

Harmful design could be seen as the outcome of end users' design practices. The concept of design practice originates primarily from professional activities aimed at creating commercial products [43]. It recognizes the ongoing challenges designers face when dealing with complex and difficult design situations [97]. Design scholarship has recognized how design practices can lead to undesirable outcomes such as dark patterns that benefit shareholder interests at the expense of end-user interests [15, 47, 117]. In addition, well-thought-out technology can also have unintended consequences [6]. Thus, the starting point of our research is to not assume the intentionality of design practices that lead to harmful design in UGVWs, but to focus on depicting extant forms of harmful design in UGVWs.

To achieve this, we used a grounded theory (GT) approach [19] to explore what constitutes harmful design in UGVWs. We chose Roblox as our study site, one of the most popular metaverse platforms, with 61.5 million daily active users as of December 2022 [18]. On Roblox, a UGVW is called an experience in the official documentation, or a game in end users' words. Our data source is the 'r/roblox' subreddit, one of the largest Roblox user communities. Following the principles of GT, we iteratively collected and analyzed Roblox users' discourses related to harmful design, resulting

DIS '23, July 10–14, 2023, Pittsburgh, PA, USA

ACM ISBN 978-1-4503-9893-0/23/07…$15.00
https://doi.org/10.1145/3563657.3595960

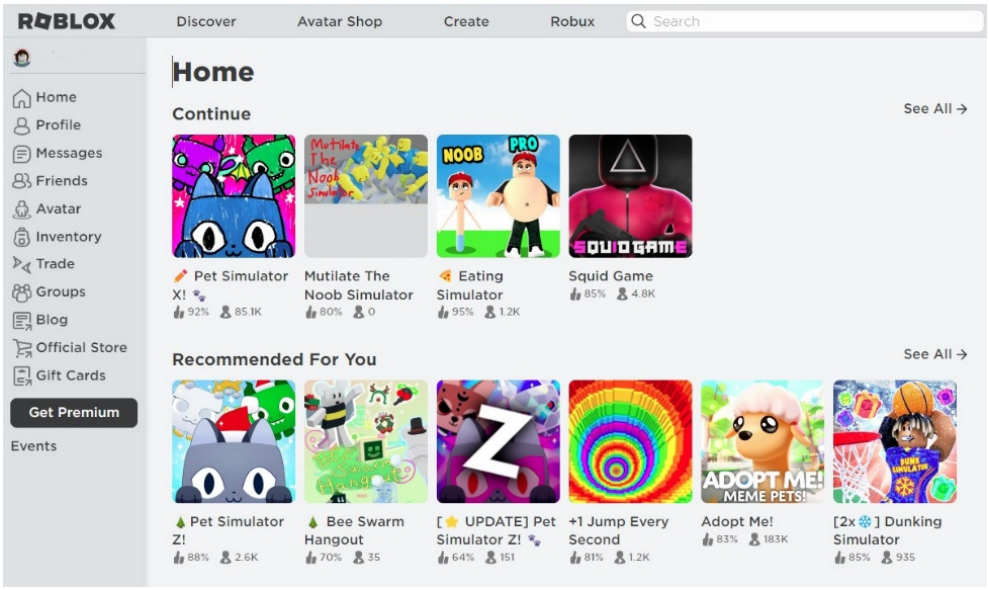

Figure 1: Home Page of Roblox.

in an explanatory framework that describes four primary types of harmful design taking place on Roblox: ubiquitous microtransaction design that nudges players to purchase in order to maximize profit, unconstrained social design that exposes players to inappropriate social interactions, unmoderated expression design that expose players to unfiltered content, and problematic world design that endorses controversial or forbidden ideologies and values. Building on this set of findings, we discuss harmful design as a distinctive form of online harm and challenges for addressing design ethics in the wild, and propose implications for design moderation (i.e., moderation mechanisms that aim to curb harmful design in UGVWs) and methods to empower child players to cope with harmful designs.

Our contributions are multi-fold: First, we present a systematic, empirical account of harmful design, contributing to critical reflection on end users' design practices. Second, we provide an analytical framework for conceptualizing the layers and components of harmful design, which lays a firm basis for future systematic intervention approaches. Third, our research is of significant broader impacts, as we explore ways to mitigate a type of online harm that targets unsuspecting users, especially child users.

2 BACKGROUND

Roblox is one of the most popular metaverse platforms, with a massive player base heavily skewed toward younger children. It had 61.5 million daily active users in December 2022, up by 18% from December 2021 [18], with the United States as the country with the most engagement time [88]. As a top online entertainment place for kids and teens [17], Roblox's primary users are under the age of 13 [103]. A 2022 report showed that 54.86% of Roblox's daily active users were under the age of 13, and 25% were under the age of 9 [28].

The Roblox homepage (see Figure 1) includes thumbnails of various UGVWs in its main block. UGVWs that the user has played in the past appear in the "Continue" category, while trending UGVWs are featured in the "Recommended For You" category. The left sidebar contains interpersonal functions such as messages and friends list, as well as purchase options through Official Store and Gift Cards. A Roblox user can visit this homepage with a web browser, and enter any UGVW. Once the user selects a UGVW on the homepage, the Roblox client (available on major operation systems such as Windows, iOS, and Android) pops up and takes the user into the UGVW.

Roblox encourages its end users to design UGVWs (which Roblox calls 'experiences' and uses call 'games') like those on the homepage, and shares revenue from the monetization of UGVWs with these UGVW creators [78]. With tens of millions of child users, roughly 5% have "published something of their own" [75]. In total, Roblox has 9.5 million developers in total, and 24 million games [81]. Roblox Studio is the development environment for end users/creators to design UGVWs, and the revenue-share model incentivizes creators to design purchasable content, such as virtual items for avatar decoration, in their UGVWs [86].

3 RELATED WORK

In this section, we draw on design and content creation scholarship to discuss the practice of creating UGVWs as a form of design practice, and then conceptualize harmful design as an ethical concern in UGVW creators' design practices.

3.1 The Creation of UGVW as Design Practice

Design practice, according to Goodman et al. [43], refers to "professional design activities intended to create commercial products." However, there are diverse notions of practice and what constitutes professional practice. To reconcile them, Goodman et al. [43] drew from Green's three core aspects [49], including activities, experiences, and contexts of practice. Thus, the very first step in

understanding and supporting design practice is to first analyze such composite of "what practitioners do, what they experience, and the context where this takes place" [43].

The attention to design practice represents critical reflections on the research-practice gap between HCI theories, methods, and pedagogy and the actual practices that design practitioners engage in [43]: there is a lack of knowledge transfer between HCI research and actual design work [89]; designers may perceive high costs in applying HCI theories or knowledge [97]; academic and industry researchers may have different perspectives on similar problems [57]; and HCI researchers may overlook the complexities and messiness in the actual design practices and treat them as scientifically solvable problems [97, 105]. Following these reflections, there has been a growing body of research that focuses on unpacking the complex situations of design practice as experienced by design practitioners that resonate with the three cores aspects of practice [49]. For example, Zhang and Wakkary [118] conducted an interview study to show how designers' personal experiences influenced their design practice, such as one designer using their personal experiences to decide what would be indispensable without doing user research. Gray et al. [46] used a content analysis of end-user discourses around problematic design patterns to discuss ethical concerns about design practices. Khemani and Reeves [59] interviewed experienced design practitioners to surface concerns about transferring design guidelines generated in HCI research to their voice user interface design practice.

While existing scholarship around design practice has concerned primarily the creations of professional designers, the lens of design practice is pertinent in analyzing the creation of UGVWs on three grounds, including creators' design agency, professionalization, and platformization. First, while much existing HCI research has cast end users as consumers of designs produced by professional designers, end users can design as well. Don Norman stated that "we are all designers" [72], since we are all constantly making conscious decisions to manipulate elements in our environment to fit with our individual lifestyles. This echoes well with the nature of design in dealing with wicked problems that do not have an absolute, clear answer, and negotiating with a messy, oftentimes constrained, solution space [97]. End users' design capabilities have been enhanced to a great degree with the advent of the Internet and its associated "participatory culture" [56] or "peer production" [7]. Along this way, the power structure between end users and designers has shifted significantly. For this point, we need to look no further than the case of video game design, which is most similar to the design of UGVW. Video game design used to take place only within professional game studios and companies [116]. However, with advances in game design tools and distribution platforms in recent decades, the technical barrier to designing and making video games has been significantly reduced, allowing more and more video game end-users to participate in video game design [30]. Much research has been done to understand two specific forms of game design created by end users, including modding [54], meaning user-generated modifications of games, and indie game design [35], which refers to game design carried out by independent designers. In both cases, end users are equipped with a range of design tools, exert design agency, and apply design thinking to create a desirable product.

Second, Goodman's conception of design practice stresses the professional quality, implying a professional/amateur dichotomy [43]. However, for design practitioners in certain fields such as UX design, there are not yet formalized procedures for determining one's eligibility or qualification to be a professional, and practitioners often rely on various self-taught paths to start a professional career [42, 44]. In addition, with the rapid diffusion of technological innovation, the line between amateur and professional is becoming increasingly blurry. For example, Smyth and Helgason [95] observed how computing technologies and physical prototyping systems, along with the rise of Do-It-Yourself culture, have blurred the boundary between amateur and professional's roles in making. Freeman et al. [35] used the term "pro-amateur" to describe how end-users could initiate technological innovation in indie game design in a bottom-up manner. On creator platforms such as YouTube, Instagram, and TikTok, many creators aspire to be professionals in producing and monetizing their creative content [92]. Thus, we refrain from taking a dichotomic view of professional practice in UGVW creation.

Third, while professional practice is often viewed as being institutionalized in organizational settings [49], especially as a profession acquires more societal recognition and occupational identity [2, 23], contemporary content creation is increasingly organized and integrated into the institutional logic of online platforms, becoming a vital part of the platform economy, a process that communications scholars refer to as platformization [25, 29]. To this end, many creator platforms such as TikTok, YouTube, and Instagram have already developed and maintained a business model that shares advertising revenue with their creators, and fosters viable career pathways for their creators [20, 68]. While individual content creators experience such an institutionalization process, they also encounter and negotiate with increased precarity and power imbalances between creators and platforms. For example, Cummingham and Craig [26] discussed how platforms exert power and control over content creators, but the latter could challenge the former by seeking alternative sources of revenue or forming collective representations. Arriagada and Ibáñez [5] found that Chilean fashion and lifestyle Instagram creators lifestyle adapt their creative practices to platform changes. Ma and Kou [64] reported how YouTube creators negotiated with the platform in order to appeal moderation decisions such as video removal.

In sum, the making of UGVW is both a form of content creation, in line with creative practices on creator platforms such as TikTok and YouTube, and a design practice, in which creators exert their design agency to effect changes in creative and complex ways. Employing the perspective of design practice, we seek to foreground complexity and messiness in designing a UGVW which are both enabled and constrained in the context of platformization, and, subsequently, to unpack the enactment of harmful design in this design practice.

3.2 Harm and Ethics in Design

Scholars have observed how harm can be embedded in design. In the classic article, *Do Artifacts Have Politics?*, Langdon Winner analyzed an example where some bridges in Long Island, New York were deliberately designed with extraordinarily low clearance to

discourage the passing of buses, thus perpetuating the designers' social-class bias and racial prejudice against people who rely on public transit [111]. The design of social institutions can be harmful as well. Broom et al. [16] questioned the notion of "unintended consequences" and pointed out how the policy design in Australia's health, welfare, and immigration systems is intended to cause harm. Along this line of inquiry, HCI researchers have also documented numerous cases where design can do harm to its end users. Apps that use manipulative interface design to gain access to users' personal data could endanger user privacy [45]. Loot boxes in video games mimic gambling mechanisms [1], resulting in not only addiction but also financial harm.

Clearly, a technological design is not necessarily a neutral artifact [12], but possesses a "script" that is intended by the designer to shape the actions of the actors [104]. To refrain from endorsing a form of technological determinism, Verbeek used the concept of mediation to further illustrate a complex interplay between users, designers, and designed technologies that co-shape the mediating role of technologies [104]. For example, if a UGVW is designed with the purpose to support interpersonal communication, it could be appropriated by users for sex acts through avatars, and the UGVW could have emergent qualities as the platform makes updates de/sexualizing avatars. Thus, design ethics concerns not just the functionality of technologies, but also their mediating roles [104].

Viewing the UGVW creation as design practice, our investigation of harmful design is informed by the value sensitive design (VSD) scholarship. Value could be broadly defined as "what a person or group of people consider important in life" [39]. Value sensitive design denotes "a theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process" [38]. VSD was motivated in the 1990s by the observation that limited attention was paid to values in computing systems [37], but researchers had already expressed concerns about key values such as privacy, autonomy, and informed consent [37, 38, 40]. Building on VSD, Flanagan and Nisenbaum [33] further proposed the values at play (VAP) framework to address values in the specific context of play. While VSD could be used to analyze values in existing computing systems, VAP is more suitable for the direct use in the game design process [33]. However, in this study, since we are exploring harmful designs in existing UGVWs other than applying values in UGVW design, both VSD and VAP can be instrumental in guiding our research.

Methodologically speaking, a complete VSD project includes three types of investigations: conceptual, empirical, and technical investigations. The conceptual investigation queries key constructs, stakeholders, and their properties and interrelationships, the empirical investigation utilizes empirical methods to contextualize technology use, while the technical investigation focuses on explicating or supporting values in system design [38]. These three types of investigations could occur in any order, although the technical investigation usually occurs after the early stage [110]. In this study, we determine that an empirical investigation is suitable for us to describe existing harmful designs in UGVWs on Roblox.

4 METHODOLOGY

The study was motivated by the gap between the sheer volume of media attention to harmful designs in UGVWs in recent years and the scarce academic attention by the time of this study. We cast harmful designs as a design issue, and, as one of the early attempts, seek to describe how harms are inserted into the design of UGVWs. Such enquiry aligns with a grounded theory (GT) methodology [24] which emphasizes theoretical construction. Pertaining to principles of GT [22, 24], the data collection and analysis processes of this study were not isolated from each other, but highly interactive and mutually constitutive. Still, we use 4.2 Data Analysis to outline key procedures and considerations specific to the data analysis process. Prior to data collection, the study received approval from the IRB office at Penn State University.

4.1 Data Collection

In line with Corbin and Strauss's suggestion that a research question can "lead researchers into the data where they can explore the issues and problems. . ." [24], we decided to use the '/r/Roblox' subreddit for our initial inquiry. The subreddit is one of the largest Roblox user communities hosted by Reddit. By the time of this study, the subreddit had more than 550 thousand subscribers. The vibrant user community and diverse discussions around Roblox allowed us to use an exploratory approach to surface end users' concerns related to harmful designs in UGVWs. End users' online discussions on social media are considered as a legitimate data source in the GT methodology [24], and have been utilized in design research (e.g., [46, 119]).

GT has several variants [22] with different emphases along dimensions such as what constitutes a resulting theory, the role of literature review, and positioning between constructivism and positivism, and the revelation of these considerations alongside the description of methodology plays a vital role in acknowledging research subjectivities and facilitating other scholars' interpretations of the results. In this regard, our approach echoes with Charmaz's GT variant [19] that emphasizes how researcher subjectivities and experiences shape not only the research process but also the outcome. Prior to this study, the researchers had years of experiences in design research and were already familiar with Roblox and lots of news reports about Roblox's harmful designs, which shaped how the researchers decided to view harmful designs as a design issue rather than just a type of online harm such as harassment and hate speech. The design angle enabled us to connect to strands of literature such as design practice and design ethics. While harmful design is a fairly open research space, we decided to focus on the types of harmful design first so as to lay a descriptive foundation for future work. The focus on types of harmful design accords with Corbin and Strauss's emphasis on developing a single theoretical category with multiple corresponding concepts [24].

Our data collection process took place in January 2023, where two researchers engaged in intensive data collection via Reddit's API, and analysis of user discussions related to harmful designs. Given that online platforms regularly update their technical functions and platform policies, we decided to focus on user discussions that took place in the year of 2022, rather than those from previous years, so that our analysis could reflect the recent state of

harmful design on Roblox. To start, we must first have a set of inclusion criteria for identifying harmful designs mentioned in user conversations. Informed by prior literature on design practices and ethics as well as rounds of discussions, we concluded that we would target user conversations that directly addressed the design or setup of a UGVW and its relationship to harm, and we would exclude conversations that only focused on harm through interpersonal interactions (e.g., one Roblox player verbally harasses another in a UGVW). A related issue was how to identify harms, especially given the uniqueness of the Roblox context which has been understudied. We approached this issue by first deriving a set of keywords from the Roblox's platform policy regarding online harm [82]. This yielded an initial set of keywords that we could use to search through the subreddit for related user conversations. The initial keywords were {'toxic,' 'moderation,' 'offensive,' 'sex,' 'predatory,' 'Nazi,' 'naked,' 'harm,' 'violent,' 'terrorist,' 'extremist,' 'bully,' 'suicide,' 'abuse,' 'illegal,' 'discriminate,' 'racist,' 'porn,' 'extort/blackmail,' 'spam,' 'harass,' and 'safe'}. For each keyword, we also searched with its variations (e.g., 'moderate' and 'moderated' for 'moderation'). Our search with these initial keywords identified a total of 1124 threads. Each of us then read through the threads to mark relevant threads, followed by discussions to resolve disagreements. This ultimately resulted in 29 threads with 2219 comments, which we found to be directly relevant to harmful design. When determining the relevance of a thread, we looked specifically for Roblox users' causal reasoning that contained both descriptions of designs as well as harms they can incur. For example, we considered a thread relevant if a Roblox user responding to it claimed that the design of the clicking simulator in UGVWs is financially exploitative towards child players. Alternatively, we considered a thread irrelevant if the post and its associated comments did not mention any instance of design or did not discuss the potential harm of a design.

We then conducted open coding on this initial dataset, an essential step in the GT methodology [22]. This step resulted in basic codes that described instances of harmful design that existed on Roblox. This initial analysis also informed our next data collection efforts, in terms of finding new keywords, to focus on identifying and theorizing data that were not previously included. This was also known as theoretical sampling in GT [24]. Through the iterative data collection and analysis, our final keyword set was {'toxic,' 'moderation,' 'offensive,' 'sex,' 'predatory,' 'Nazi,' 'naked,' 'harm,' 'violent,' 'terrorist,' 'extremist,' 'bully,' 'suicide,' 'abuse,' 'illegal,' 'discriminate,' 'racist,' 'porn,' 'extort/blackmail,' 'spam,' 'harass,' 'safe,' 'monetization,' 'gambling,' 'worry,' 'troll,' 'policy,' 'ban,' 'delete,' 'remove,' 'da hood,' 'condo games,' 'Russian Roulette,' 'traumatic,' 'not suitable for work (NSFW),' 'dangerous,' 'terms of service (TOS),' 'manipulate,' 'parent,' 'child,' 'son,' 'daughter,' 'inappropriate,' 'wrong,' 'underground,' 'bad game,' and 'bad design'}. Our final dataset included 54 threads with 4618 comments. At this point, we decided that we reached "theoretical saturation" [13], where no new ideas were found. Through the whole data collection and analysis processes, the researchers constantly took memo to record emergent ideas and linkages between data and ideas [24].

4.2 Data Analysis

Open coding is the process of "breaking data apart and delineating concepts to stand for interpreted meaning of raw data" [24]. A code tends to use informal language to describe what is in the raw data. We performed open coding continuously upon the initial 29 threads as well as all the new data that were added to form the final dataset. For example, a comment that described how a UGVW contained numerous designs that tricked players to spend money through microtransactions would be coded as "greedy design patterns intended for cash grab." Our analysis resulted in a total of 71 initial codes.

We then conducted axial coding on our initial codes as well as their associated data, a process that relates initial codes through a combination of inductive and deductive thinking [24]. To be specific, this was an iterative process. Each of the two researchers analyzed the list of initial codes, grouping codes with similar meanings. For example, one initial code was "core gameplay mechanism that was unmoderated" and described a spray paint simulator which players could use to express abusive ideas, and another initial code was "inappropriate avatar appearance," referring to how players could decorate their avatars with offensive symbols. These two initial codes could be associated as both described instances of expression design that enabled harm. The two researchers then held discussions to resolve disagreements. This allowed us to generate four major types of harmful design that form a single theoretical category of harmful design patterns, including:

Ubiquitous microtransaction design describes the ubiquity of microtransactions implemented in UGVWs, potentially resulting in financial harm to UGVW players.

Unconstrained social design refers to the phenomenon that social spaces in UGVWs are designed without taking the presence of children into consideration, potentially exposing them to inappropriate or harmful content.

Unmoderated expression design describes various communication mechanisms that UGVW players can use to express ideas without proper moderation.

Problematic world design captures the storytelling of a UGVW's general setting that pertains to controversial or even extremist ideologies.

After completing the axial coding process, we realized that our thought process and reasoning were influenced by several underlying rationales, whether we were aware of them at the time of coding or not. First, we sought to categorize those design instances in terms of their affordances, or how a design could enable possible actions. For example, a deceptive microtransaction design could trigger reckless spending, while an unconstrained social design such as a hidden social space might encourage sexually explicit interactions. Second, our action of categorization was influenced by how we perceived the type of harm. Most evidently, the ubiquitous microtransaction design leads to economic harm, unconstrained social design and unmoderated expression design expose players to inappropriate content, while problematic world design often embodies concepts of radicalization and extremism. Lastly, our thought process also included a desire to order the categories based on their level of abstraction. Clearly, microtransaction designs are often specific and straightforward to understand, but what constitutes a

world design is more abstract and contains multiple interconnected components.

In the next section, we report our analytical results. We have paraphrased all the quotes when reporting to reduce their searchability while preserving their original meaning.

5 FINDINGS

In this section, we describe four primary harmful design patterns we surfaced from Roblox users' online discussions, including ubiquitous microtransaction design that induces financial harm, unrestricted social design that enables inappropriate interpersonal interactions, unmoderated expression design without proper moderation, as well as problematic world design where the overall story perpetuates harmful ideologies.

5.1 Ubiquitous Microtransaction Design

Microtransaction refers to a mechanism that allows players to purchase in-game virtual items with money. Microtransaction has been a common business model in online games released by professional studios [100]. For each item sold on the Roblox marketplace, the revenue-share model awards 30% of the total amount to the creator, 40% to the seller or distributor, and 30% to the platform [87]. However, players cannot purchase these virtual items directly through real money. Instead, they must first use real money to purchase Robux, the virtual currency in Roblox (one Robux is approximately 0.01U.S. Dollars). However, the rate for creators to cash out their earned Robux is significantly lower, at "$0.0035U.S. Dollars per Earned Robux" [84]. Thus, it is unsurprising that Roblox has been criticized for exploiting young creators [76] and capitalizing on child labor [75].

Currently, there is currently no regulation on microtransaction in online games, and online games only practice self-regulation [66]. Subsequently, microtransactions in UGVWs are even more unregulated, leading to various ethical concerns within the Roblox user community.

One common ethical concern is that experienced creators can quickly design and release low-quality UGVWs for the sake of a "cash grab," meaning a product designed primarily to generate money. To achieve this, they would copy and paste the source code of their existing UGVWs, altering only a few elements to qualify as a new one. For example, a Roblox user observed:

> Within an hour of the movie "turning red" being released, there were tens of platformers under that theme. Every one of them was filled to the brim with microtransactions. These are just cash grabs. The creators used random content to make a game and title it with something related to the movie, as well as a clickbait thumbnail of that thing. When a player plays it, they get a pop-up window every ten seconds asking for Robux to speed up. They are just trying to get kids to pay for these items for a one-time use. The creators are just taking advantage of the fact that the players are too young to understand that their items are way overpriced.

'Platformer' is a game genre in which a player controls their character to move from one platform to another (e.g., Super Mario games). The quote above describes how Roblox creators could follow trends in popular culture, such as a newly released movie, in their design ideation of UGVWs. Roblox hosts millions of UGVWs and more are added every day. These UGVWs compete for players' attention and money. Subsequently, UGVW creators are financially incentivized to incorporate popular trends into their designs in order to attract Roblox players. Experienced creators could reuse the source code from their previous UGVWs for a different theme and create a newly themed UGVW by making only a few cosmetic changes (such as changing the background color to match the new theme). Since the goal is to maximize profits during a temporary period following a trend, creators may not prioritize the quality and reputation of the UGVW, nor plan for its long-term maintenance. However, unsuspecting players, such as a fan of the said movie, "turning red," could be attracted to such types of UGVWs and be willing to spend a large sum of money through deliberate microtransaction design patterns such as a "pop-up window" that interrupts the player experience for monetary gain. Thus, the user expressed worries about how such UGVW design is specifically tailored to cause financial harm to young players.

Not only are microtransactions easy to implement in UGVWs, but they also include gambling mechanisms, further encouraging players' addictive and reckless spending behavior. Another user commented:

> Loot boxes are rampant in the most popular Roblox games like Adopt Me and Murder Mystery. This is gambling for children who want to get rarer items than others. This is the same trick used in other online games, and loot boxes are ruining people's lives.

Loot box is a type of microtransaction where a player pays money to acquire a random item. The rarer the item, the less chance the player could get it. Thus, loot boxes are usually associated with gambling [63, 69], and already outlawed in several countries [41, 99]. However, Roblox is based in the U.S., which has no regulation on loot boxes. Thus, the user was concerned that popular UGVWs on Roblox were already exploiting such microtransaction designs to induce financial harm to child players.

Roblox users further questioned Roblox's role in not only tolerating such designs, but intentionally encouraging them. A player wrote:

> Every Roblox game has gambling in the form of loot boxes. Also, the limited items in the avatar catalog are obviously gambling. Everyone on Roblox is doing this, and this is how Roblox makes money. They do not get into trouble because gambling laws don't apply to video games, even though there is in fact real money involved.

Here, the user explained that such microtransaction design could bring enormous economic benefits to Roblox as a for-profit corporation, and that Roblox has skillfully navigated the legal system to sustain its profitable business model. What makes this microtransaction design particularly concerning is the scale at which Roblox operates and the potential player population it affects. As a result, the business model is working as intended, at the expense of the financial wellbeing of its players.

5.2 Unconstrained Social Design

Social design means design considerations around creating a social space where players can socialize with each other. However, online social design must be taken carefully when children are involved. For example, unlike other social platforms, Roblox's policy clearly states that "*Roblox is a safe space for meeting online friends, chatting, and collaborating on creative projects, but we prohibit content that seeks or portrays romantic relationships, including: Animations of kissing, hand holding, or other romantic gestures in a romantic context; Experiences that depict romantic events, including weddings, dates, and honeymoons*" [82]. However, Roblox users have reported encountering explicit content or interactions due to the harmful social design of UGVWs, typically in one of two forms. First, some social spaces are designed with the intent to cause interpersonal harm. For example, one particular type of UGVW is called "condo game," where players could enter a condo-like social space and encounter sexual acts and explicit language. Several Roblox users discussed the problems with "Condo games" in the following conversation:

> R1: Condo games were made to be inappropriate.
>
> R2: What are condo games?
>
> R3: They are literally porn games and get taken down within a few minutes.
>
> R4: Not always. They have workarounds and use new accounts, so they are less noticeable. Players must have their discord to join. It won't be banned unless it gets leaked.

In the conversation above, Roblox users discussed condo games as an example of unconstrained social design. Particularly, some creators would knowingly design inappropriate UGVWs like this, and bypass the moderation system by making the UGVW invite-only. While condo games are a distinctive type of UGVW, it is far more common that players enter an unknown UGVW and encounter harm. Towards this end, a Roblox user recommended that:

> Just stay away from any games that have a lot of players named as slenders. Most are online daters and some even do NSFW [short for 'not suitable for work'] games.

Roblox allows users to design and customize their avatars, and slender on Roblox refers to a specific avatar apperarnce that is thin and tall [94]. Online dating is not allowed on Roblox due to its predominantly young user base. In the quote above, the user observed that certain social designs attracted slenders, and slenders tended to possess malicious intents. Thus, the user stated that social designs populated by slenders were more likely to be risky and inappropriate and should be avoided if possible.

Second, some social spaces are vulnerable and prone to abuse and become a hotbed for interpersonal harm. Two Roblox users conversed:

> R1: I joined one of those "vibe place hangout games" and what I saw was really distrubing. A lot of people were using emotes to have "sex." In addition, people were exchanging Tiktoks, Snapchat and Kik. This is very disturbing because the kids could easily fall victim to predators, groomers, and kidnappers.
>
> R2: I also spent time in one of these games and what you said is true. People were trying to share their discord tags, typing very inappropriate things on roblox by bypassing the filter with weird characters and spamming x's at the beginning of words, role-playing as couples having intercourse by walking in and out of girls repeatedly, using very obscure euphemisms like "beating the meat," and trying to ask for people's personal information.

In the conversation excerpt, R1 and R2 shared their similar experiences with inappropriate interpersonal interactions, ranging from sex acts and crude language that are strictly forbidden by Roblox policy, to information solicitation which could be risky for child players. Although hangout games are not like condo games, they do provide a venue where inappropriate interpersonal interactions could easily occur. However, there are limited design considerations to make such social venues safer. To this point, a Roblox user observed that "*The online daters have turned the party feature of MeepCity into nothing but porn clubs.*"

5.3 Unmoderated Expression Design

Expression design addresses a variety of channels designed within UGVWs where players can express their ideas. While designing a channel for free expression appears innocuous, it becomes problematic when the channel is embedded in a UGVW and left unmoderated. Unmoderated channels can be filled with inappropriate content. Specifically, our analysis uncovered three types of expression that are enabled by the unmoderated expression design. First, some UGVWs' core gameplay mechanisms give players tools to express ideas. Two Roblox users discussed this issue:

> R1: The Roblox community is ridiculously toxic and rampant with homophobia, racism, and misogyny. You can check how prevalent this is in games that allow users to express their opinions freely (Rate my avatar, spray paint games, military games etc) ... I find it surprising that parents are okay with their children playing Roblox, given the amount of inappropriate games and the toxic community in general.
>
> R2: Spray Paint gets a lot of hate because it is extremely unmoderated. You can basically write whatever you want in Spray Paint. This is also true in Free Draw 2.

The Roblox users pointed to specific UGVWs with creative tools such as spray painting and drawing that give players the power of free expression but do not have any proper oversight. As a result, such unmoderated expression design leaves much room for harmful user-generated content such as "homophobia, racism, and misogyny."

Second, some UGVWs' overall setting may have designated channels for players' expressions. These channels also lack moderation and are vulnerable to problematic expressions. Regarding this, two Roblox users conversed:

> R1: I was playing a game called Booth Plaza, and I saw a booth with child porn for the booth's image... This is not uncommon. Every day or two there is a

> booth with a guy decapitated or guts spewing out. This really crossed the line.
>
> R2: . . . Booth games is a breeding ground for that type of stuff because it allows anyone to type anything they want in a booth and with a decal on it. . . Roblox tries to ban them, but people evolve with Roblox's moderation. . .
>
> R1: Booth Plaza mods should have intervened. But none of them cares. They only join to make jokes and powerstrip to feel special.
>
> R2: Yeah, booth games are fun to socialize with other players, but terrible because people can put anything they want in a booth and it will be public. Decals are unfiltered once they accidentally bypass the system.

In the conversation above, the Roblox users talked about a UGVW designed as a virtual plaza where a player could have a virtual booth and then decorate it with their own expressions, in forms such as image, decal, or language. While Roblox's platform-level moderation system does general screening of any content uploaded by a UGVW player, the system could be bypassed, as R2 indicated. Thus, the task of moderating player expressions currently is delegated to each specific UGVW's moderation. The approach is ineffective, as R1 observed. As a result, designated channels within a UGVW, such as a virtual booth on a virtual plaza, could become a place for inappropriate content.

Lastly, players could also make free expressions through their avatars' appearances, including virtual clothing. Below are observations made by several Roblox users:

> I have seen on Roblox that a person in a UGVW was wearing a Nazi armband.
>
> The game [Clear Skies over Milwaukee] got me scared because the players had questionable outfits.
>
> My child was in MeePCity and ran into a character wearing a t-shirt with an image of sex act on it. He was traumatized.

The three examples above came from different UGVWs, but they all point to the same issue with unmoderated expression design — players could freely express any content through their avatars' appearance, regardless of the appropriateness and risk the content carries.

5.4 Problematic World Design

World design refers to the setting where all stories happen, and its characters develop. What makes a virtual world unique is that it is different from the real world that we live in and has "its own unique aesthetic feel and flavor" [114]. Each UGVW has its own world design that provides a backstory for players to interact with, and some world designs are deeply problematic in endorsing controversial and even extremist ideas and ideologies.

World designs could make representations of violence and abuse highly graphic. For example, a few Roblox users discussed an incident they encountered:

> R1: I played Criminality today, and was immediately sold as a maid. Is this common in this type of games?
>
> R2: What the hell? You were sold as a maid? What is wrong with this game? Roblox has a lot of weird things, but I have never heard of player trafficking on Roblox. What happened exactly?
>
> R1: They just knocked down my character and then carried it to another player, who then paid some money.
>
> R2: This is so wrong.
>
> R3: Criminality has an underground-ish audience, so I'm not surprised.

In the excerpt above, Criminality is a popular UGVW on Roblox in which players freely roam and fight each other. Underground refers to a design practice on Roblox that hides inappropriate content under the surface to avoid moderation. R1 described a violent experience in the said UGVW and sought explanation from the user community. Based on R1's description, R2 and R3 both echoed with such experience while pointing out the ethical problem in such design. In a similar vein, our analysis also identified instances where UGVWs simulate mass school shootings or suicidal bombings.

World designs could also reproduce and perpetuate real-world stereotypes. Several Roblox users complained about this design issue:

> R1: All the "hood games" where kids posing as gangsters and bypassing community, etc. They are very problematic.
>
> R2: I don't like these games, because they have this idea that this is what black people are like and this is what black neighborhoods are like. It's horrible. . . It is offensive as hell. I don't like these creators who think they know what the hood is like. It's all those white kids pretending to be gangsters. It's awful.

Hood games are a type of UGVWs that make an underclass neighborhood setting where players act as gangsters to engage in various criminal activities. In the conversation above, R1 and R2 acutely found issue with such world design that tends to employ racial stereotypes against African Americans, as well as prejudices against lower socio-economic classes.

Sometimes, world designs can embody prejudices in an implicit way. For example, another Roblox user noticed:

> R1: I've entered places such as Papers, Please, which was once rebranded as a generic Soviet government thing. They are well made, but the people there are really horrible. I suspect that the military aspect attracts people with certain views. When I was there, I saw people making jokes that suggest questionable beliefs. . . They seem to hold prejudices against minorities or are really nationalistic to a weird degree. That's just the feeling I get.
>
> R2: Power-hungry children want to boss people around. This is a toxic hierarchy that is enforced in many military role-playing games.

According to the observations by R1 and R2, UGVWs that feature military roleplaying mechanisms could embody values such as hierarchy, prejudice, and nationalism, resonating with values commonly celebrated in the real-world military culture [96]. As

such, R1 expressed ethical concerns that UGVWs of such world design could foster online groups that harbor extremist thoughts.

Lastly, Roblox users also brought up world designs that represent harmful ideologies. For example, a Roblox user mentioned:

> Many years ago, I had a lot of time to spend so I found a game. It was a simple game, where two teams fought to win gold and buy better equipment... One day I was playing the game as normal and was defeated. Then, I was respawned in another room full of WW2 German flags. I wasn't there alone. I tried to leave the room but couldn't, so I just left the game.

In the above quote, the Roblox user described a world design with inappropriate content that is associated with Nazi. What appeared to the user at the beginning, as a simple game, only revealed later the kind of ideology it tried to promote. Although the UGVW was probably removed by Roblox moderation, the impact of the harmful design remained on the Roblox user and perhaps other players who were "respawned" in the same room.

6 DISCUSSION

We have surfaced four primary harmful design patterns in the creation of user-generated virtual world (UGVW) on Roblox. These harmful designs exist at different conceptual levels and originate from the empowered capabilities for end users to create complex, dynamic content on metaverse platforms like Roblox. The harmful designs are not mutually exclusive and can co-exist in a single UGVW. Given the complexity of the harm we described, it is particularly suitable to use design as an analytic angle. Next, we discuss how harmful design differs from previously discussed harm types in light of design practice, reflect on design ethics in the wild, and propose implications for mitigating it through design moderation and better supporting child players.

6.1 The Complexity of Harmful Design

Harmful design in UGVWs is categorically more complex than commonly discussed online harm in user-generated content (UGC). The very forms of UGC (e.g., text, audio, image, or user behavior) define parameters for how online harm is conveyed and causes negative effects. For example, hate speech in text could be read by target groups and hurt them mentally and emotionally [115]; pictures depicting violence could be viewed by children and have a disturbing effect on their minds [93]; and harmful user behaviors such as harassment are often enabled by particular platform affordances such as virtual avatars in social virtual reality [36] and interpreted as harmful by victims. In these examples, UGC causes harm through a cognitive process in which a harmful meaning is captured, perceived, and interpreted by victims. However, harmful designs, as our findings suggest, could cause harm in a myriad of ways beyond the communication of harmful meanings: they could manipulate players into purchasing behavior through ubiquitous microtransaction design, disguise inappropriate interpersonal interactions in unconstrained social design, expose players to unfiltered information and content via unmoderated expression design, and dispose players to problematic or extremist ideas, values, and ideologies through problematic world design. The harm from these designs is only more severe when players are unwitting and unsuspecting children, a victim group which Roblox users frequently mentioned in our findings.

To unpack the complexity of harmful design, we could contemplate several directions: design resources, design space, and design knowledge. First, UGVW creators have access to *a wider variety of design resources* than UGC creators. They can maneuver not only static content such as text, image, video, and audio, but also algorithmic objects and procedures to produce a desired effect. For example, microtransaction design relies upon algorithmic thinking to determine when and how to seek money from players; and world design utilizes an assemblage of algorithms to orchestrate a toxic hierarchical experience. Although there has been a great deal of effort in technical approaches that focus on filtering and autodetecting unwanted usage to prevent risks for users (e.g., [21, 70]). Such approaches primarily deal with textual content using machine learning and natural language processing techniques including supervised learning, lexicon-based, rule-based, and mixed-initiative approaches [91]. Although these approaches are useful to some extent, they suffer from a lack of quality datasets, limitations in capturing contextual information (e.g., participants' relationship), and capturing all criteria of unsafe content [4, 90, 91]. In the case of Roblox, it has a chat filtering system that filters out inappropriate language or other unsafe textual content such as swear words, personally identifiable information, and words associated with bullying and harassment [85], and automatically checks whether avatars are "wearing appropriate attire within the avatar editor and avatar thumbnails" [83]. However, because such technical approaches focused primarily on static contents, they cannot address the design issues within UGVWs, where harms are in real-time, interactive, and diverse form. Overall, while UGVW platforms could be efficient at detecting and removing harmful static content, they are less effective in moderating algorithmically and procedurally induced harms, which would require human interpretation to determine and mitigate the existence of harms.

Second, UGVW creators are exploring *a widely open design space*, compared to the creative space a video content creator on YouTube or an image creator on Instagram could explore. Design addresses a "wicked problem" [79] which does not have a definitive or final solution. Creators have numerous ways to approach their design objective — they can make an avatar-mediated social space, simulate real-world scenarios such as military roleplay, develop video games, etc.

Third, UGVW creators are developing *highly situated design knowledge*, pertaining to the Roblox platform. That is, they observe, acquire, and evolve design knowledge that could help achieve their design objective, whether to attract more players or to make more money. Designers could draw on their knowledge to anticipate how users interact with their technological design [104], creators would also anticipate how players interact with their UGVWs and designate particular mechanisms such as microtransactions to take advantage of such anticipation, or virtual booths to facilitate inappropriate communication.

Prior research has highlighted how questionable designs stem from the practices of professional designers, such as dark patterns that benefit shareholders and the expense of end users' interests [15, 47] and profit-driven microtransactions in video game design

[69]. These conversations presume a dichotomy between designer and end user, and power imbalance between them. What we showed in this study is how questionable designs could also stem from end users and inflict harm on other end users. Technology democratizes not only who can use and access it, but also who can generate harm. While professional designers are financially incentivized to produce questionable designs, UGVW creators' harmful design practices are shaped by both the profit-driven platform economy and poor platform governance.

6.2 Design Ethics of End Users in the Wild

In light of the research-practice gap [73], prior academic research has developed high-level ethical frameworks such as VSD [37] and VAP [33] but also looked into ethical concerns regarding professional designers' work [11, 45]. Extending this line of concern, what we surfaced in this work is ethical concerns regarding end users' work. To be more accurate, it is not all the end users that participate in designing UGVWs and producing harmful designs. UGVW creators are a specific end-user groups who assume a 'designer' role and acquire a more important role than other end users. Vilaza et al. [74] observed how ethics research in HCI tends to focus less on specific end-user groups and more on technology types, tacitly assuming all individuals equally experience technology. In this study, it is clear to us the so-called Roblox users are not a homogeneous group, and the end users play out a diverse range of roles and capabilities to engage with the Roblox platform.

When creators are equipped with design tools to solve a design problem in the wild, they are also facing questions (from the broader Roblox user community as we revealed in the findings) about design ethics that commonly face professional designers. It is with no doubt that design ethics is something important to consider in the context of UGVW creation. While much attention has been paid to the ethics of professional designers through both scholarly scrutiny and industry self-regulation [48, 67], such attention has not yet existed for UGVW creators. As a result, the only normative boundaries for UGVW creators seem to be just what Roblox's moderation detects and deletes.

This is not to say that UGVW creators lack ethical agency — the free will and capability to make ethical design decisions. According to van der Velden, agency is "not solely in the hands in the user but in particular socio-material configurations of designers, technology, and users" [102]. In this view, the (un)ethical agency to make harmful designs are not solely in the hands of UGVW creators, but dynamically configured within the interactions between UGVW creators, the Roblox platform, as well as UGVW players. This is most evident in "virtual plaza" UGVWs, where the creators knowingly create and maintain a free expression venue, the Roblox platform imposes little oversight, and (some of) the players start to abuse the unmoderated expression design. In this example, the ethical agency is distributed across multiple technologies and actors, and through their interactions.

The distributed nature of ethical agency in UGVW creation calls for a multi-stakeholder framework to enhancing design ethics. Our findings identified multiple stakeholder roles, such as UGVW platform owner, UGVW creators, UGVW players, and UGVW players' parents. Recent years' extensive media reports (e.g., [14, 98]) also suggested external entities that care about UGVWs' ethical issues. These stakeholders can acquire different ethical responsibilities in collectively defining and enforcing ethical practices. With the increasing platformization of UGVW creation, the platform should recognize that their organizational and institutional relationship with creators and guide their ethical practice, through means such as developing ethics training program [107]. UGVW creators should recognize that designing a UGVW requires not only technical expertise but also ethical awareness and expertise in determining what is or is not appropriate. UGVW players, who could already recognize ethical violations in UGVWs, could play a more active role in debating and enforcing ethical values in UGVW creation. Both UGVW players' guardians and external entities should also be involved in assessing the appropriateness of UGVW design for children.

6.3 From Content Moderation to Design Moderation

Our study outlined four primary types of harmful design. Following GT principles [24], the resulting types came from not only what Roblox users perceived but also how the researchers interpreted user perceptions. In light of the observation that specific, detailed categories which may misrecognize or exclude certain harm types [10], our classification of harmful design types is best read in a generative sense, aimed at opening a research space for broader discussions of harmful design initiated by end users. Also because of the heavy involvement of human interpretation in determining the harmfulness of a design, the means to moderate harmful design will be decidedly different from content moderation that targets UGC that unfolds well within platform-defined parameters, such as text, image, audio, and user behavior. Here, for the purpose of this discussion, user behavior could be considered as a type of UGC.

Existing moderation approach (e.g., ones that Roblox is using) falls short in regulating harmful designs because it originates from moderating UGC instead of design practice [50, 65, 90, 91], with the assumption that harm already exists in the content to be found, and technical features could be extracted to detect the harm (e.g., regular expressions to predict offensive phrases). However, harmful designs, as we discussed in this study, can be dynamic and emergent in players' on-the-fly interactions with UGVWs. Certain types of UGVWs, such as "condo games" or military roleplaying games, are prone to abuse and exploitation. However, the examples of unmoderated expression design show that adverse players could invent novel ways to be harmful when certain design permits. Thus, we need an updated understanding of moderation tailored for the case of user-generated harmful design.

Design moderation, in this context, is defined as a set of governance strategies to foster benign UGVW designs while inhibiting harmful design patterns. Design moderation must pertain to the distinctive characteristics of harmful designs to be effective in enforcing normative boundaries. Since harmful designs are dynamic and emergent, they are discernible to end users but elusive for automated moderation techniques. Thus, design moderation should consider how to leverage the ethical expertise of end users, so that human knowledge could be integrated in the decision-making

process of moderation. Participatory governance is a pertinent perspective here. Prior moderation and governance scholarship has explored various ways, such as allowing users to develop policies and institutional procedures in Wikipedia [34] and giving users technical means to upvote or downvote as distributed moderation on Slashdot [61]. While Roblox currently relies on AI-enforced moderation, UGVW platforms in general could leverage participatory governance and empower end users in more important moderation roles. For example, Roblox users could be invited to collectively make community guidance on what constitutes harmful design. Roblox users could also play a more important role in adjudicating problematic UGVW designs reported by other end users.

6.4 Increasing Children's Awareness of and Resilience to UGVW Harm

Given that many UGVW platforms have children as their primary user population, it is critical to raise children's awareness of and foster their resilience to harmful designs in UGVWs. Prior literature on children's online safety has proposed various methods, such as designing control methods for risk prevention, such as age limits and parental control [52]. However, age limits could be easily manipulated [51], and hardly reflect the complex landscape of millions of UGVWs on Roblox. Parental control could lead to overprotection of children, neglect of children's autonomy, distrust between parents and children, and hinder developmental processes that are critical to teach children how to protect themselves [51]. Researchers have suggested to seek a balance between parental control and children's agency and self-regulation [112]. In this regard, a viable path could be to provide opportunities for parents and their children to explore UGVWs together. Relatedly, in our data collection process, in our iterative expansion of search keywords, we added new terms such as 'daughter' and 'son' to describe parent-child relationships. This was done in response to encountering several instances where a Roblox user claimed to be playing with their children to explore the vast array of UGVWs on Roblox. This way, parents and children play more equal and collaborative roles, where parental protection is at work implicitly.

In addition to focusing on risk prevention and "viewing children's online safety as something that depends on the actions of others" [60], it is also critical to explore how to promote children's resilience so that they can effectively and wisely protect themselves from online risks [113]. Online safety education is a critical way to empower children to use the internet in safe and responsible ways [3, 106]. There have been an increasing number of online safety educational programs developed and marketed by national and international actors [32], such as guidebooks for parents (e.g., [31, 109]) and teachers (e.g., [53]), and educational games for children (e.g., [8, 9]). Thus, education programs and materials can be developed to focus on UGVWs. For example, educational materials could be developed dedicated to describing typical problematic world designs and illustrating their harms to parents and teachers, so that they can enhance their understanding of the potential risks that children may encounter while using a platform that appears to be child-friendly.

6.5 Limitations

Our data was collected from conversations that happened within one Roblox user online community during a given period of time. Thus, the specific instances of harmful design that we identified may not be directly generalizable to other metaverse platforms or other time periods. The conceptual categories are meant to be read in a generative fashion, inspiring future empirical endeavors to expand our understanding of the scale and nature of harmful design in the metaverse. We acknowledge how researcher subjectivities impacted our methodological thinking and choices, and believe that such acknowledgement lends strength to the contextualization, validity, and transparency of our study results.

7 CONCLUSION

This study explored the issue of harmful design on Roblox, a metaverse platform. By explicating four types of harmful designs, we outline novel ways end users engage in design practice and generate harms targeted at a primarily child population. We cast harmful design as both a design ethics issue and a moderation issue and discuss potential solutions. Much more work is needed in the future to explore, measure, and mitigate harmful design on metaverse platforms that give end users powerful tools to design and create novel forms of harm.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their cogent and constructive feedback. The work is partly supported by NSF No. 2006854.

REFERENCES

[1] Jacob Aagaard, Miria Emma Clausen Knudsen, Per Bækgaard, and Kevin Doherty. 2022. A Game of Dark Patterns: Designing Healthy, Highly-Engaging Mobile Games. *Conference on Human Factors in Computing Systems - Proceedings* (April 2022). DOI:https://doi.org/10.1145/3491101.3519837

[2] Andrew Abbott. 1991. The Order of Professionalization: An Empirical Analysis. *Work Occup* 18, 4 (November 1991), 355–384. DOI:https://doi.org/10.1177/0730888491018004001

[3] Zainab Agha, Zinan Zhang, Oluwatomisin Obajemu, Luke Shirley, and Pamela J. Wisniewski. 2022. A Case Study on User Experience Bootcamps with Teens to Co-Design Real-Time Online Safety Interventions. *Conference on Human Factors in Computing Systems - Proceedings* (April 2022). DOI:https://doi.org/10.1145/3491101.3503563

[4] Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. *SIGIR 2019 - Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (July 2019), 45–53. DOI:https://doi.org/10.1145/3331184.3331262

[5] Arturo Arriagada and Francisco Ibáñez. 2020. "You Need At Least One Picture Daily, if Not, You're Dead": Content Creators and Platform Evolution in the Social Media Ecology. *Social Media and Society* 6, 3 (July 2020). DOI:https://doi.org/10.1177/2056305120944624/ASSET/IMAGES/LARGE/10.1177_2056305120944624-FIG2.JPEG

[6] Joan S. Ash, Marc Berg, and Enrico Coiera. 2003. Some Unintended Consequences of Information Technology in Health Care: The Nature of Patient Care Information System-related Errors. *Journal of the American Medical Informatics Association* 11, 2 (November 2003), 104–112. DOI:https://doi.org/10.1197/jamia.M1471

[7] Yochai Benkler. 2006. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale University Press. Retrieved September 28, 2014 from http://books.google.com/books/about/The_Wealth_of_Networks.html?id$=$X0Q0mAEACAAJ&pgis$=$1

[8] Berkman Klein Center For Internet & Society. 2016. The Internet and You.

[9] Livio Bioglio, Sara Capecchi, Federico Peiretti, Dennis Sayed, Antonella Torasso, and Ruggero G. Pensa. 2019. A Social Network Simulation Game to Raise Awareness of Privacy among School Children. *IEEE Transactions on Learning Technologies* 12, 4 (October 2019), 456–469. DOI:https://doi.org/10.1109/TLT.2018.2881193

[10] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *Proc ACM Hum Comput Interact* 1, CSCW (2017), 24. DOI:https://doi.org/10.1145/3134659

[11] Kerstin Bongard-Blanchy, Arianna Rossi, Salvador Rivas, Sophie Doublet, Vincent Koenig, and Gabriele Lenzini. 2021. I am Definitely Manipulated, even When i am Aware of it. It's Ridiculous! - Dark Patterns from the End-User Perspective. *DIS 2021 - Proceedings of the 2021 ACM Designing Interactive Systems Conference: Nowhere and Everywhere* (June 2021), 763–776. DOI:https://doi.org/10.1145/3461778.3462086

[12] Albert Borgmann. 1987. *Technology and the Character of Contemporary Life: A Philosophical Inquiry*. University of Chicago Press.

[13] Glenn A. Bowen. 2008. Naturalistic inquiry and the saturation concept: a research note. *Qualitative Research* 8, 1 (February 2008), 137–152. DOI:https://doi.org/10.1177/1468794107085301

[14] Russell Brandom. 2021. Roblox is struggling to moderate re-creations of mass shootings. *The Verge*. Retrieved from https://www.theverge.com/2021/8/17/22628624/roblox-moderation-trust-and-safety-terrorist-content-christchurch

[15] Harry Brignull. 2011. Dark Patterns: Deception vs. Honesty in UI Design. *Interaction Design, Usability*, 338.

[16] Alex Broom, Michelle Peterie, Katherine Kenny, Gaby Ramia, and Nadine Ehlers. 2022. The administration of harm: From unintended consequences to harm by design. *Crit Soc Policy* (April 2022). DOI:https://doi.org/10.1177/02610183221087333

[17] Business Wire. 2018. Roblox Emerges as a Top Online Entertainment Platform for Kids and Teens in 2017 - Roblox. *Bloomberg*. Retrieved from https://www.bloomberg.com/press-releases/2018-03-21/roblox-emerges-as-a-top-online-entertainment-platform-for-kids-and-teens-in-2017

[18] Ashley Capoot. 2023. Roblox stock up after December update shows increase in bookings. *CNBC*. Retrieved February 2, 2023 from https://www.cnbc.com/2023/01/17/roblox-stock-up-after-december-update-shows-increase-in-bookings.html

[19] Kathy Charmaz. 2006. *Constructing grounded theory: a practical guide through qualitative analysis*. Sage Publications.

[20] Clement Chau. 2010. YouTube as a participatory culture. *New Dir Youth Dev* 2010, 128 (December 2010), 65–74. DOI:https://doi.org/10.1002/YD.376

[21] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. *Proceedings - 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust and 2012 ASE/IEEE International Conference on Social Computing, SocialCom/PASSAT 2012* (2012), 71–80. DOI:https://doi.org/10.1109/SOCIALCOM-PASSAT.2012.55

[22] Tom Cole and Marco Gillies. 2022. More than a bit of coding: (un-)Grounded (non-)Theory in HCI. *Conference on Human Factors in Computing Systems - Proceedings* (April 2022). DOI:https://doi.org/10.1145/3491101.3516392

[23] Peter Conrad and Joseph W. Schneider. 2011. Professionalization, monopoly, and the structure of medical practices. In *The Sociology of Health and Illness*. 156–162.

[24] Juliet M. Corbin and Anselm L. Strauss. 2015. *Basics of qualitative research: techniques and procedures for developing grounded theory* (4th. ed.). SAGE Publications, Inc.

[25] Stuart Cunningham and David Craig. 2019. Creator Governance in Social Media Entertainment. *Soc Media Soc* 5, 4 (November 2019). DOI:https://doi.org/10.1177/2056305119883428

[26] Stuart Cunningham and David Randolph Craig. 2019. *Social media entertainment: the new intersection of Hollywood and Silicon Valley*. Retrieved January 14, 2023 from https://books.google.com/books/about/Social_Media_Entertainment.html?id$=$FQ2BDwAAQBAJ

[27] Cecilia D'Anastasio. 2021. How "Roblox" Became a Playground for Virtual Fascists | WIRED. *WIRED*. Retrieved from https://www.wired.com/story/roblox-online-games-irl-fascism-roman-empire/

[28] Brian Dean. 2022. Roblox User and Growth Stats 2022. *backlinko*. Retrieved from https://backlinko.com/roblox-users

[29] Brooke Erin Duffy, Thomas Poell, and David B. Nieborg. 2019. Platform Practices in the Cultural Industries: Creativity, Labor, and Citizenship: https://doi.org/10.1177/2056305119879672 5, 4 (November 2019). DOI:https://doi.org/10.1177/2056305119879672

[30] Sean C. Duncan. 2010. Gamers as Designers: A Framework for Investigating Design in Gaming Affinity Spaces. *E-Learning and Digital Media* 7, 1 (January 2010), 21–34. DOI:https://doi.org/10.2304/ELEA.2010.7.1.21

[31] Shawn Marie. Edgington. 2011. The parent's guide to texting, facebook, and social media: understanding the benefits and dangers of parenting in a digital world. (2011), 188.

[32] David Finkelhor, Kerryann Walsh, Lisa Jones, Kimberly Mitchell, and Anne Collier. 2021. Youth Internet Safety Education: Aligning Programs With the Evidence Base. *Trauma Violence Abuse* 22, 5 (December 2021), 1233–1247. DOI:https://doi.org/10.1177/1524838020916257

[33] Mary Flanagan, Daniel C Howe, and Helen Nissenbaum. 2005. Values at Play: Design Tradeoffs in Socially-Oriented Game Design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2005). DOI:https://doi.org/10.1145/1054972

[34] Andrea Forte, Vanesa Larco, and Amy Bruckman. 2009. Decentralization in Wikipedia Governance. *J. Manage. Inf. Syst.* 26, 1 (July 2009), 49–72. DOI:https://doi.org/10.2753/MIS0742-1222260103

[35] Guo Freeman, Nathan McNeese, Jeffrey Bardzell, and Shaowen Bardzell. 2020. "Pro-Amateur"-Driven Technological Innovation: Participation and Challenges in Indie Game Development. *Proc ACM Hum Comput Interact* 4, GROUP (January 2020). DOI:https://doi.org/10.1145/3375184

[36] Guo Freeman, Samaneh Zamanifard, Divine Maloney, and Dane Acena. 2022. Disturbing the Peace: Experiencing and Mitigating Emerging Harassment in Social Virtual Reality. *Proc ACM Hum Comput Interact* 6, CSCW1 (April 2022). DOI:https://doi.org/10.1145/3512932

[37] Batya Friedman. 1996. Value-sensitive design. *interactions* 3, 6 (December 1996), 16–23. DOI:https://doi.org/10.1145/242485.242493

[38] Batya Friedman, Peter H. Kahn Jr., and Alan Borning. 2002. *Value Sensitive Design: Theory and Methods*.

[39] Batya Friedman, Peter H. Kahn, Alan Borning, and Alina Huldtgren. 2013. Value Sensitive Design and Information Systems. *Philosophy of Engineering and Technology* 16, (2013), 55–95. DOI:https://doi.org/10.1007/978-94-007-7844-3_4/FIGURES/5

[40] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Trans Inf Syst* 14, 3 (July 1996), 330–347. DOI:https://doi.org/10.1145/230538.230561

[41] Tom Gerken. 2018. Video game loot boxes declared illegal under Belgium gambling laws. *BBC News*. Retrieved February 10, 2023 from https://www.bbc.com/news/technology-43906306

[42] Guiseppe Getto, Liza Potts, Michael J. Salvo, and Kathie Gossett. 2013. Teaching UX: Designing programs to train the next generation of UX experts. *SIGDOC 2013 - Proceedings of the 31st ACM International Conference on Design of Communication* (2013), 65–69. DOI:https://doi.org/10.1145/2507065.2507082

[43] Elizabeth Goodman, Erik Stolterman, and Ron Wakkary. 2011. Understanding interaction design practices. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, ACM Press, New York, New York, USA, 1061–1070. DOI:https://doi.org/10.1145/1978942.1979100

[44] Colin M. Gray. 2016. "It's More of a Mindset Than a Method": UX Practitioners' Conception of Design Methods. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, ACM Press, New York, USA, 4044–4055. DOI:https://doi.org/10.1145/2858036.2858410

[45] Colin M. Gray, Jingle Chen, Shruthi Sai Chivukula, and Liyang Qu. 2021. End User Accounts of Dark Patterns as Felt Manipulation. *Proc ACM Hum Comput Interact* 5, CSCW2 (October 2021). DOI:https://doi.org/10.1145/3479516

[46] Colin M. Gray, Shruthi Sai Chivukula, and Ahreum Lee. 2020. What kind of work do "asshole designers" create? Describing properties of ethical concern on reddit. *DIS 2020 - Proceedings of the 2020 ACM Designing Interactive Systems Conference* (July 2020), 61–73. DOI:https://doi.org/10.1145/3357236.3395486

[47] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. 2018. The Dark (Patterns) Side of UX Design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, ACM Press, New York, New York, USA, 1–14. DOI:https://doi.org/10.1145/3173574.3174108

[48] Colin M Gray and Shruthi Sai Chivukula. 2019. Ethical Mediation in UX Practice. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019). DOI:https://doi.org/10.1145/3290605

[49] William Green. 2009. Introduction: Understanding and Researching Professional Practice. In *Understanding and Researching Professional Practice*. Sense Publishers, 1–18. Retrieved February 3, 2023 from https://researchoutput.csu.edu.au/en/publications/introduction-understanding-and-researching-professional-practice

[50] Wan Noor Hamiza Wan Ali, Masnizah Mohd, and Fariza Fauzi. 2019. Cyberbullying Detection: An Overview. *Proceedings of the 2018 Cyber Resilience Conference, CRC 2018* (January 2019). DOI:https://doi.org/10.1109/CR.2018.8626869

[51] Heidi Hartikainen, Netta Iivari, and Marianne Kinnula. 2016. Should We design for control, trust or involvement? A discourses survey about children's online safety. *Proceedings of IDC 2016 - The 15th International Conference on Interaction Design and Children* (June 2016), 367–378. DOI:https://doi.org/10.1145/2930674.2930680

[52] Yasmeen Hashish, Andrea Bunt, and James E. Young. 2014. Involving children in content control: A collaborative and education-oriented content filtering approach. *Conference on Human Factors in Computing Systems - Proceedings* (2014), 1797–1806. DOI:https://doi.org/10.1145/2556288.2557128

[53] Sameer Hinduja and Justin W. Patchin. 2009. Bullying beyond the schoolyard: preventing and responding to cyberbullying. (2009), 254.

[54] Renyi Hong. 2013. Game Modding, Prosumerism and Neoliberal Labor Practices. *International Journal of Communication 7*, 19. Retrieved September 14, 2020 from https://ijoc.org/index.php/ijoc/article/view/1659

[55] Julie Jargon. 2021. Roblox Struggles With Sexual Content. It Hopes a Ratings System Will Address the Problem. *The Wall Street Journal*. Retrieved from https://www.wsj.com/articles/roblox-struggles-with-sexual-content-it-hopes-a-ratings-system-will-address-the-problem-11618660801

[56] Henry Jenkins. 2006. *Convergence Culture: Where Old and New Media Collide*. NYU Press. DOI:https://doi.org/10.3395/reciis.v2i1.165pt

[57] Arnowitz Jonathan and Elizabeth Dykstra-Erickson. 2005. CHI and the practitioner dilemma. *Interactions*, 5–9. Retrieved February 3, 2023 from https://dl.acm.org/doi/fullHtml/10.1145/1070960.1070964?casa_token$=$Y6njzJgFLQIAAAAA:9DbDwX1rCwED-gFPjaus3VuhOoh4kMOlZdLXLTS83XayI8Xmt38TNULkxxkEtwKrxzCJEbqPglpN7w

[58] Sean Keach. 2018. Roblox kids' game haven for Jihadi, Nazi and KKK roleplay featuring Twin Tower bombings and race-hate murders. *The Sun*.

[59] Krishika Haresh Khemani and Stuart Reeves. 2022. Unpacking Practitioners' Attitudes Towards Codifications of Design Knowledge for Voice User Interfaces. *Conference on Human Factors in Computing Systems - Proceedings* (April 2022). DOI:https://doi.org/10.1145/3491102.3517623

[60] Priya Kumar, Shalmali Milind Naik, Utkarsha Ramesh Devkar, Marshini Chetty, Tamara L. Clegg, and Jessica Vitak. 2017. "No Telling Passcodes Out Because They're Private": Understanding Children's Mental Models of Privacy and Security Online. *Proc ACM Hum Comput Interact* 1, CSCW (December 2017). DOI:https://doi.org/10.1145/3134699

[61] Cliff Lampe and Paul Resnick. 2004. Slash(dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the 2004 conference on Human factors in computing systems - CHI '04*, ACM Press, New York, New York, USA, 543–550. DOI:https://doi.org/10.1145/985692.985761

[62] Stuart Lauchlan. 2022. Game on! Microsoft's near $70 billion gambit on the metaverse. *diginomica*. Retrieved from https://diginomica.com/game-microsofts-near-70-billion-gambit-metaverse

[63] Kevin Liu. 2019. A Global Analysis into Loot Boxes: Is It Virtually Gambling. *Washington International Law Journal* 28, 3 (2019), 763–800.

[64] Renkai Ma and Yubo Kou. 2022. "I am not a YouTuber who can make whatever video I want. I have to keep appeasing algorithms": Bureaucracy of Creator Moderation on YouTube. In *CSCW'22 Companion: Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing*, Association for Computing Machinery (ACM), 8–13. DOI:https://doi.org/10.1145/3500868.3559445

[65] Sean MacAvaney, Hao Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PLoS One* 14, 8 (August 2019), e0221152. DOI:https://doi.org/10.1371/JOURNAL.PONE.0221152

[66] Kishan Mistry. 2018. P(L)aying to Win: Loot Boxes, Microtransaction Monetization, and a Proposal for Self-Regulation in the Video Game Industry. *Rutgers University Law Review* 71, (2018). Retrieved February 10, 2023 from https://heinonline.org/HOL/Page?handle$=$hein.journals/rutlr71&id$=$547&div$=$11&collection$=$journals

[67] Stephanie Moore. 2014. Ethics and design: Rethinking professional ethics as part of the design domain. *Design in Educational Technology: Design Thinking, Design Process, and the Design Studio* (January 2014), 185–204. DOI:https://doi.org/10.1007/978-3-319-00927-8_11/TABLES/1

[68] J. Morreale. 2014. From homemade to store bought: Annoying Orange and the professionalization of YouTube. *Journal of Consumer Culture* 14, 1 (March 2014), 113–128. DOI:https://doi.org/10.1177/1469540513505608

[69] Erica L. Neely. 2019. Come for the Game, Stay for the Cash Grab: The Ethics of Loot Boxes, Microtransactions, and Freemium Games. https://doi.org/10.1177/1555412019887658 16, 2 (November 2019), 228–247. DOI:https://doi.org/10.1177/1555412019887658

[70] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. *25th International World Wide Web Conference, WWW 2016* (2016), 145–153. DOI:https://doi.org/10.1145/2872427.2883062

[71] Josh Norem. 2022. Meta CEO Says It Expects One Billion People to Be in The Metaverse by 2030. *Extreme Tech*. Retrieved from https://www.extremetech.com/internet/337425-meta-ceo-says-it-expects-one-billion-people-to-be-in-the-metaverse-by-20230

[72] Donald A. Norman. 2004. Epilogue: We Are All Designers. In *Emotional Design*. Basic Books.

[73] Donald A. Norman. 2010. The research-practice gap. *Interactions* 17, 4 (July 2010), 9–12. DOI:https://doi.org/10.1145/1806491.1806494

[74] Giovanna Nunes Vilaza, Kevin Doherty, Darragh McCashin, David Coyle, Jakob Bardram, and Marguerite Barry. 2022. A Scoping Review of Ethics Across SIGCHI. *DIS 2022 - Proceedings of the 2022 ACM Designing Interactive Systems Conference: Digital Wellbeing* 18, (June 2022), 137–154. DOI:https://doi.org/10.1145/3532106.3533511

[75] Simon Parkin. 2022. The trouble with Roblox, the video game empire built on child labour. *The Guardian*. Retrieved from https://www.theguardian.com/games/2022/jan/09/the-trouble-with-roblox-the-video-game-empire-built-on-child-labour

[76] People Make Games. Investigation: How Roblox Is Exploiting Young Game Developers. Retrieved January 12, 2023 from https://www.youtube.com/watch?v$=$_gXlauRB1EQ

[77] Felix Pope. 2022. Children's game Roblox features Nazi death camps and Holocaust imagery. *The Jewish Chronicle*. Retrieved from https://www.thejc.com/news/news/childrens-game-roblox-features-nazi-death-camps-and-holocaust-imagery-128DrQ3MoW1jzQa17oym2S

[78] Amanda Reaume. 2022. How Does Roblox Make Money? *Seeking Alpha*. Retrieved from https://seekingalpha.com/article/4486523-how-does-roblox-make-money

[79] Horst W.J. Rittel and Melvin M. Webber. 1973. Dilemmas in a general theory of planning. *Policy Sci* 4, 2 (June 1973), 155–169. DOI:https://doi.org/10.1007/BF01405730/METRICS

[80] Andy Robertson. 2022. Parents guide to Roblox and how your kids can play it safely. *Internet Matters*. Retrieved from https://www.internetmatters.org/hub/esafety-news/parents-guide-to-roblox-and-how-your-kids-can-play-it-safely/#:$\sim$:text$=$After all%2C much of Roblox,with little supervision and understanding.

[81] Roblox. Home - Roblox.

[82] Roblox. 2022. Roblox Community Standards. *Roblox.com*. Retrieved from https://en.help.roblox.com/hc/en-us/articles/203313410-Roblox-Community-Standards

[83] Roblox. 2022. For Parents.

[84] Roblox. 2023. Developer Exchange Terms of Use. *Roblox Support*. Retrieved February 2, 2023 from https://en.help.roblox.com/hc/en-us/articles/115005718246-Developer-Exchange-Terms-of-Use

[85] Roblox. Safety Features: Chat, Privacy & Filtering.

[86] Roblox Creator Documentation. Luau. Retrieved January 12, 2023 from https://create.roblox.com/docs/scripting/luau

[87] Roblox Creator Documentation. Developer Economics. Retrieved January 12, 2023 from https://create.roblox.com/docs/production/monetization/economics

[88] Robolox. 2022. A YEAR ON ROBLOX: 2021 IN DATA.

[89] Yvonne Rogers. 2004. New theoretical approaches for human-computer interaction. *Annual Review of Information Science and Technology* 38, 1 (September 2004), 87–143. DOI:https://doi.org/10.1002/aris.1440380103

[90] H. Rosa, N. Pereira, R. Ribeiro, P. C. Ferreira, J. P. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A. M. Veiga Simão, and I. Trancoso. 2019. Automatic cyberbullying detection: A systematic review. *Comput Human Behav* 93, (April 2019), 333–345. DOI:https://doi.org/10.1016/J.CHB.2018.12.021

[91] Semiu Salawu, Yulan He, and Joanna Lumsden. 2020. Approaches to Automated Detection of Cyberbullying: A Survey. *IEEE Trans Affect Comput* 11, 1 (January 2020), 3–24. DOI:https://doi.org/10.1109/TAFFC.2017.2761757

[92] Ryan Schram. 2020. The state of the creator economy: Ingenta Connect. *Journal of Brand Strategy* 9, 2 (2020), 152–162. Retrieved August 10, 2022 from https://www.ingentaconnect.com/content/hsp/jbs/2020/00000009/00000002/art00007

[93] Ariana Shahinfar, Nathan A. Fox, and Lewis A. Leavitt. 2000. Preschool Children's Exposure to Violence: Relation of Behavior Problems to Parent and Child Reports. *American Journal of Orthopsychiatry* 70, 1 (January 2000), 115–125. DOI:https://doi.org/10.1037/H0087690

[94] Udayveer Singh. 2022. What is a Roblox Slender and Who Created It? | Beebom. *Beebom*. Retrieved February 10, 2023 from https://beebom.com/what-is-roblox-slender/

[95] Michael Smyth and Ingi Helgason. 2017. Making and Unfinishedness: Designing Toolkits for Negotiation. *The Design Journal* 20, sup1 (July 2017), S3966–S3974. DOI:https://doi.org/10.1080/14606925.2017.1352899

[96] Joseph L. Soeters, Donna J. Winslow, and Alise Weibull. 2006. Military Culture. *Handbooks of Sociology and Social Research* (2006), 237–254. DOI:https://doi.org/10.1007/0-387-34576-0_14/COVER

[97] Erik Stolterman. 2008. The Nature of Design Practice and Implications for Interaction Design Research. *International Journal of Design* 2, 1 (2008), 55–65.

[98] Rachel Stonehouse. 2019. Roblox: "I thought he was playing an innocent game." *BBC News*. Retrieved from https://www.bbc.com/news/technology-48450604

[99] Nicholas Straub. 2020. Every Country With Laws Against Loot Boxes (& What The Rules Are). *ScreenRant*. Retrieved February 10, 2023 from https://screenrant.com/lootbox-gambling-microtransactions-illegal-japan-china-belgium-netherlands/

[100] The Business Research Company. 2022. *Online Microtransaction Market Analysis, Size And Trends Global Forecast To 2022-2030*. Retrieved February 10, 2023 from https://www.thebusinessresearchcompany.com/press-release/online-microtransaction-market-2022

[101] Joe Tidy. 2021. Zuckerberg's metaverse: Lessons from Second Life - BBC News. *BBC News*. Retrieved January 14, 2023 from https://www.bbc.com/news/technology-59180273

[102] Maja van der Velden. 2009. Design for a common world: On ethical agency and cognitive justice. *Ethics Inf Technol* 11, 1 (December 2009), 37–47. DOI:https://doi.org/10.1007/S10676-008-9178-2/METRICS

[103] VentureBeat. Roblox believes user-generated content will bring us the Metaverse. *2020*.

[104] Peter Paul Verbeek. 2016. Materializing Morality: Design Ethics and Technological Mediation. *Sci Technol Human Values* 31, 3 (August 2016), 361–380. DOI:https://doi.org/10.1177/0162243905285847

[105] Ron Wakkary. 2006. Framing complexity, design and experience: A reflective analysis. *Digital Creativity* 16, 2 (2006), 65–78. DOI:https://doi.org/10.1080/14626260500173013

[106] Kerryann Walsh, Elizabeth Pink, Natasha Ayling, Annette Sondergeld, Elizabeth Dallaston, Paul Tournas, Ella Serry, Sharon Trotter, Tia Spanos, and Nada Rogic. 2022. Best Practice Framework for Online Safety Education: Results from a rapid review of the international literature, expert review, and stakeholder consultation. *Int J Child Comput Interact* 33, (September 2022), 100474. DOI:https://doi.org/10.1016/J.IJCCI.2022.100474

[107] Danielle E. Warren, Joseph P. Gaspar, and William S. Laufer. 2014. Is Formal Ethics Training Merely Cosmetic? A Study of Ethics Training and Ethical Organizational Culture. *Business Ethics Quarterly* 24, 1 (January 2014), 85–117. DOI:https://doi.org/10.5840/BEQ2014233

[108] Anna Werner. 2020. Kids exposed to simulated sex and graphic images in "dark side" of popular computer game Roblox - CBS News. *CBS News*. Retrieved February 2, 2023 from https://www.cbsnews.com/news/roblox-condo-games-kids-exposed-pornographic-scenes-sex-acts/

[109] Pamela Whitby. 2011. Is your child safe online?: a parent's guide to the internet, Facebook, mobile phones & other new media. White Ladder.

[110] Till Winkler and Sarah Spiekermann. 2021. Twenty years of value sensitive design: a review of methodological practices in VSD projects. *Ethics Inf Technol* 23, 1 (March 2021), 17–21. DOI:https://doi.org/10.1007/S10676-018-9476-2/FIGURES/1

[111] Langdon Winner. 1980. Do Artifacts Have Politics? *Modern Technology: Problem or Opportunity* 109, 1 (1980), 121–136.

[112] Pamela Wisniewski, Arup Kumar Ghosh, Heng Xu, Mary Beth Rosson, and John M Carroll. 2017. Parental Control vs. Teen Self-Regulation: Is there a middle ground for mobile online safety? *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (2017). DOI:https://doi.org/10.1145/2998181

[113] Pamela Wisniewski, Haiyan Jia, Na Wang, Saijing Zheng, Heng Xu, Mary Beth Rosson, and John M Carroll. 2015. Resilience Mitigates the Negative Effects of Adolescent Internet Addiction and Online Risk Exposure. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15), ACM, New York, NY, USA, 4029–4038. DOI:https://doi.org/10.1145/2702123.2702240

[114] Mark J.P. Wolf. 2017. World Design. *The Routledge Companion to Imaginary Worlds* (January 2017), 67–73. DOI:https://doi.org/10.4324/9781315637525-9/WORLD-DESIGN-MARK-WOLF

[115] MichałWypych and MichałBilewicz. 2022. Psychological toll of hate speech: The role of acculturation stress in the effects of exposure to ethnic slurs on mental health among Ukrainian immigrants in Poland. *Cultur Divers Ethnic Minor Psychol* (2022). DOI:https://doi.org/10.1037/CDP0000522

[116] Christopher James Young. 2018. Game Changers: Everyday Gamemakers and the Development of the Video Game Industry. University of Toronto.

[117] José P Zagal, Staffan Björk, and Chris Lewis. 2013. Dark Patterns in the Design of Games. In *Foundations of Digital Games*. Retrieved September 11, 2022 from http://urn.kb.se/resolve?urn$=$urn:nbn:se:ri:diva-24252

[118] Xiao Zhang and Ron Wakkary. 2014. Understanding the role of designers' personal experiences in interaction design practice. In *Proceedings of the 2014 conference on Designing interactive systems - DIS '14*, ACM Press, New York, New York, USA, 895–904. DOI:https://doi.org/10.1145/2598510.2598556

[119] Annuska Zolyomi. 2021. Where the stakeholders are: tapping into social media during value-sensitive design research. *Ethics Inf Technol* 23, 1 (March 2021), 59–62. DOI:https://doi.org/10.1007/S10676-018-9475-3/METRICS