Taxonomy of Interaction Techniques to Mitigate Inappropriate Social Interactions in Social Virtual Reality

Arthur Audrain, Katja Zibrek, Ferran Argelaguet

Taxonomy of Interaction Techniques to Mitigate Inappropriate Social Interactions in Social Virtual Reality

Arthur Audrain, Katja Zibrek, and Ferran Argelaguet

Abstract—Social Virtual Reality (SVR) enables users to embody avatars and interact with others in shared virtual environments. On platforms such as Rec Room, VRChat, and Meta Horizon Worlds, inappropriate social behaviours are common and can lead to harmful emotional experiences for users. To prevent or mitigate these negative effects, SVR platforms offer a range of actions, referred to as safety tools, that allow users to control the information and interactions they are exposed to. Although prior research has analysed and proposed initial classifications of these tools, a structured, theory-driven approach to their characterization is still lacking. In this paper, we propose a taxonomy of safety tools for SVR platforms inspired by existing literature in social psychology and human–computer interaction. In order to characterise these tools, we focus on the area of social interaction, which occurs between a sender and a receiver, within a social context of the SVR medium. Our approach not only provides a model of social interaction in SVR, but also describes how safety tools function, including the processes of selecting and manipulating them, factors that are critical to their effectiveness. Finally, we apply our framework to characterise existing safety tools, identify their limitations, and present design guidelines for future SVR platforms.

Index Terms—Social Virtual Reality, safety tools, inappropriate behaviour.

1 Introduction

Social Virtual Reality (SVR) is a VR application where users can engage in immersive and embodied social interactions with others. Platforms such as Rec Room [3], VRChat [5] and Meta Horizon Worlds [1] are the most prominent examples. Compared to traditional social network platforms, social interactions in SVR happen in real time and aim at replicating real social interactions through the use of VR technology, such as full body tracking, avatars, and multisensory rendering [29,40]. Through VR technology, SVR enables users to freely interact with other users [33,34], fosters creativity in expression through their avatars [22,25] and supports customisable content, and thus provides new and exciting ways for people to meet and interact [19]. Similarly as in physical reality, however, the SVR platforms can be places where inappropriate behaviour takes place, which can potentially cause emotional and even physical harm to the user [11,15,24,41]. The question of what is inappropriate behaviour in SVR is complex, as it strongly depends on the individual perception of each user as to which behaviours they deem inappropriate [43]. Users of SVR tend to define it as any behaviour or interaction that intentionally generates discomfort, goes against their will, and/or causes interpersonal harm [21]. These behaviours can range from hateful speeches and simulation of violence, to being annoying or disagreeing on a sensible topic.

To tackle the problem of inappropriate behaviours, SVR platforms have developed several features: for the most extreme risks (e.g., racism, simulation of sexual assault, hate speech), they have forbidden these behaviours in their code of conduct and have implemented moderation systems to punish the culprits and to discourage such behaviours; but for interactions that are more ambiguous, the most common strategy is to provide tools (i.e., interaction techniques) that allow users to manipulate/control social interactions and to adjust the SVR experience to their preferences, i.e., *safety tools*. However, past studies have shown that these tools have a number of limitations: they excessively limit the interaction possibilities [14], lack in customisation options for the user [14,43], can put too much pressure on users and be complex to use [7,51], can lack in effectiveness against certain risks [51,53] or be misused as weapons [13,14,16]. Furthermore, users might not even be aware that safety tools exist [7,51,53].

The research field of safety methods in Social VR is recent, and studies tend to focus on exploration and description of new findings, rather than developing theoretical frameworks to characterise these methods, and this lack of theoretical grounding could undermine the studies in terms of their rigour, reusability, and societal relevance of their conclusions [52]. Moreover, because SVR is a relatively new social media platform, most of the safety feature designs have been imported from other social media (e.g., mute, block and report features), without considering the particularities of the SVR medium, such as the possibility of multi-sensory interaction or the embodiment of avatars in first-person view. These limitations can be explained due to the lack of knowledge and guidelines for SVR developers on these safety features [7]. This has led to inconsistent terminology; for example, the safety feature *Block*, common in social media, does not work in the same way in different SVR platforms: in *VRChat* [5], for example, the avatar of the user targeted with an inappropriate behaviour can be made completely invisible to the other user and vice versa, but in *Rec Room* [3], the avatar of the target can only turn semi-transparent and not vice versa. Hereinafter, we define safety tools, a term used in the literature but not precisely defined, as: interaction techniques that modulate social interactions between multiple actors, with the purpose of avoiding or mitigating the impact of inappropriate social interactions.

In this paper, we present a taxonomy of safety tools in SVR following the method proposed by Nickerson et al. [36] for the creation of taxonomies. To our knowledge, this is the first taxonomy created in this area that follows a structured method. We first define the goal of the taxonomy: a decomposition of the safety methods for researchers and designers of SVR applications. Then, we used existing research to conceptualise the studied area. This involved iterating between empirical research on safety tools and existing characterisations in order to conceptualise the safety tools for SVR. The existing characterisations were inspired by the task analysis and decomposition models from the 3DUI research, specifically the work of Bowman et al. [12] as well as existing models of social interactions in social psychology [26,38]. We also put emphasis on the task decomposition of safety tools, where we make important distinctions between different steps in the activation process of the safety tools. To complete our work on the taxonomy, we provide some examples of how existing safety tools can be classified based on our framework. The paper concludes with a discussion on how and why the creation of such a taxonomy provides the required

- *Arthur Audrain is with Inria, Univ Rennes, CNRS, IRISA. E-mail: arthur.audrain@inria.fr.*
- *Katja Zibrek is with Inria, Univ Rennes, CNRS, IRISA. E-mail: katja.zibrek@inria.fr.*
- *Ferran Argelaguet is with Inria, Univ Rennes, CNRS, IRISA. E-mail: ferran.argelaguet@inria.fr.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

formalism and structure to identify key aspects of future research and give some possible directions for the design of safety tools.

2 Related Works

2.1 Inappropriate social interactions in SVR

Precisely defining if a social interaction is inappropriate is challenging, as it can depend on the individual perception of the interaction [43]. For example, users of SVR tend to define inappropriate social interactions as any behaviours or interactions that intentionally discomfort them, go against their will, and cause interpersonal harm [21] and/or are non-consensual [43]. In the context of online video games, such behaviours have been labelled as Toxic Behaviors [27], and defined as "an umbrella term used to describe various types of negative behaviours, including harassment, flaming, trolling (e.g., gaining enjoyment from intentionally annoying other players), and cheating during games". A cause of the inappropriate behaviours observed in Social Virtual Reality could be the *Gaming Culture*, which was frequently described as toxic, sexist, and intolerant [7, 8, 44, 51]. However, while these behaviours can be intentional, they can also be due to incomprehension and non-compatible desires [21]. Moreover, inappropriate behaviours can also be targeting specific/under-represented target groups such as women [24,30,37,41,42], the LGBTQ community [20,23,24,30,42] or ethnic minorities [24,30,42]), which face additional risks. Finally, it is important to notice that an inappropriate social behaviour that starts in SVR could continue outside, in other types of social media, which leads to cyber-harassment and heavier consequences for the victim [24].

Even though risks observed in non-VR online environments (e.g., Second Life, Minecraft) can also occur in social virtual reality (e.g., cyberbullying through text-messages, insults in oral communication channel), the possibility of pseudo-realistic interaction allowed by the VR medium makes inappropriate behaviours in SVR different from any other form of media [24]. For example, the personal space of users can be invaded because of the capacity of users to embody avatars and move freely, and full-body animation enables the use of offensive gestures [24]. Besides, the embodiment permitted by the VR medium intensifies the impact of these inappropriate behaviours compared to traditional online games [11, 41]. In the following we summarise past studies that provided classifications tackling inappropriate social behaviours in SVR.

First, from the testimonies of users who were victims of harassment in SVR, Blackwell et al. [11] defined three types of harassment (inappropriate behaviours): (1) Verbal Harassment: such as personal insults or hate speech, (2) Physical Harassment: such as unwanted touch on the avatar or the obstruction of user movements, and (3) Environmental Harassment: such as displaying sexual or violent content or throwing objects. This classification has been used in recent studies [18,45], for a simulation that covers all forms of toxic behaviours. This classification is based on the modalities used to convey the inappropriate behaviour, such as the voice, the body movements or the virtual environment.

Second, Zheng et al. [53] classified the safety risks in SVR, which can be considered as inappropriate social behaviours, in 8 categories, based on the **content** of the behaviour. These categories have been identified by an analysis of YouTube videos, showing risks in SVR. These forms of toxic behaviors are (1) Emerging safety risks (risks unique to VR), (2) Virtual Violence (physical attack on the victim's avatar), (3) Virtual Crashing (using strategies and bugs to ruin others' experience), (4) Virtual Scarring (using a scary-looking avatar, running towards someone...), (5) Virtual Abuse (attacks based on gender, ethnicity...), (6) Virtual Sexual Harassment (sexual assaults, sexually provocative gestures...), (7) Virtual Voice-Trolling (using a gender-mismatch voice to mock the other), and (8) Virtual Trash Action (detrimental actions to others that do not fall in one of the previous categories).

Finally, from a review of existing literature, Weerasinghe et al. [51] identified 3 categories of unsafe SVR situations (caused by inappropriate behaviors), based on how the situation is perceived by the victim (auditorily, visually or "physically"): (1) Verbal situations (inappropriate language or unwanted conversations), (2) Visual situations (exposure to offensive or disturbing visual content), and (3) Physical situations (physical assault, unwelcome touch).

2.2 Safety Tools

To tackle the issue of inappropriate social interactions, the SVR platforms (e.g., VRChat, Rec Room, and Meta Horizon Worlds) propose several interaction techniques that can be used to restrict the social interactions between two or more users. Based on observations and user studies, previous studies have classified them, examined their use, and presented guidelines for their improvement.

2.2.1 Classifications of Safety Tools

Past studies [51,53] explored the platform and the methods they have implemented to mitigate toxic behaviours and classified these techniques into several categories describing their use, primarily into three major ones, based on when they are supposed to be used: (1) Boundary Settings, before the toxic interaction; (2) Quick Reactions, after the toxic interaction; and (3) Agreements, always active and the basis of the moderation system. These two classifications are similar and presented in detail in Table 1.

These classifications were based on functionalities of existing safety tools. These functionalities can be found across platforms and have similarities in the way they are used. This allows researchers to compare the accessibility of safety tools between platforms [51], how users are using them to face harassment and how they perceive them [51], and to provide general guidelines for their design [53]. Therefore, this type of classification shows the differences of safety tools that belong in the same category, and prevent studies on the use of a safety tool category between platforms. For example, the safety tool *Block* available in RecRoom [3] makes the targeted user become semi-transparent (so still perceptible to the user), whereas the one in VRChat [5] makes the targeted user completely invisible to the user. Another example is the Personal Space Bubble (i.e., *Intimacy Proxemics*) of RecRoom, which is customizable. In contrast, the same technique in VRChat is not customizable, which can be perceived by users of being too restrictive, thus leading them to disable it.

2.2.2 Existing guidelines and propositions for safety tools

In addition to the works analysing existing safety tools, there are a number of works that have explicitly explored new interaction techniques to face inappropriate behaviours.

Because an inappropriate behaviour can be seen as the violation of a user's consent by another, a number of studies have used the lens of Consent to think of new ways to enable users to protect themselves. By working in several participatory workshops with users of VR dating apps, Zytko et al. [54] proposed *Consent Mechanics*, features that allowed users to clearly express their consent for social interactions. They proposed three types of consent mechanics: (1) giving consent to interaction, ensuring the consent when interacting with another user before it occurs through multi-stage consent bubbles; (2) informing consent to interaction, such as providing information that affects the decision to give consent to interact with users through identity verification; and (3) informing consent for observation of other users, such as providing information that affects the decision to give consent to enter a virtual environment through a tag system for virtual worlds.

In a user study, Schulenberg et al. [43] highlighted that *Consent Mechanics* already exist in the platforms but with limitations. They proposed three principles: to take inspiration from existing and used tools, to make them dynamic and customisable, and to allow users to revoke given consent at any time.

Other works proposed their own system of *Consent Mechanics*. Liao et al. [28] designed a system of *Preference Badges*, that allows users to display visually their social preferences (their current level of social energy, if they accept physical contact or not, and if they want a quiet environment) to indicate to others what kind of interaction they consent to. This system of badges is linked to an *Intimacy Proxemic*, so when the user decides to wear the "No physical contact" badge, a bubble will surround them that blocks any intruder. Wang et al. [49] created a system of *Negotiated User-to-User Teleportation*, which is a type of *Intimacy Proxemics*, which allows users to protect their personal space from the teleportation of intruders, but at the same time, they can negotiate the entry to some part with others.

Table 1: Summary of the classification of Safety tools proposed by Zheng et al. [53] and Weerasinghe et al. [51].

Meta categories	Zheng et al. [53]	Weerasinghe et al. [51]	Definition
Boundary settings	Intimacy Proxemics	Intimacy Proxemics	Allow users to establish and maintain a distance with others in the virtual environment.
	Social Spaces	Social Spaces	Allow users to create and define specific virtual areas as safe spaces.
		Content Gating	Allow users to control who can see and interact with certain types of content.
	Avatar Shields	Interaction Shields	Allow users to customise the control of their avatars when interacting with others.
	Trust Reputation + Interactions shields	Intimacy Rank	"Trust Rank" features to allow users the customisation of their experience based on the trust level of other users.
	User Demographics	Parental Control	Allow parents/tutors to impose limits for underage users to manage/limit their SVR experience.
	Voice control		Allow users to control how they perceive other voices and how their voice is perceived by others.
Quick Reaction	Safety Reactions	Safety Gestures	Allow users to activate safety features using hand movements, focusing on quick and easy access.
	Safety Zone	Safe Zone Teleport	Allow users to instantly teleport themselves to a designated Safe Zone, to block all interactions with others.
		Freeze Controls	Allow users to temporarily freeze the control of their avatars to prevent other users from engaging.
	Vote Kick	Vote Kick	Allow users to initiate a vote to remove another user from a specific virtual space.
		Mute and Block	Allow users to silence specific individuals and/or to prevent them from temporarily or permanently interacting with them.
	Safety Reports	Report	Allow users to flag inappropriate or harmful behaviour to moderators or platform administrators.
Agreements	Codes of Conduct		Description of the actions/behaviours that are appropriate or not in the SVR environment. Must be accepted before joining.
	Informed Consent		Description of safety and privacy risks that users can face in the SVR platform. Must be accepted before joining.
Other		AI Moderation	The use of AI methods (automatic) to monitor user interactions, content, and behaviour to identify potential violations of community guidelines and safety concerns.

After conducting a study on how users tend to react in a toxic situation in social VR, Zheng et al. [53] reported that they often respond with natural reactions (similar than in reality) when they are stressed, even if it is not effective. For example, while being exposed to loud sound or verbal harassment, some users try to cover their ears with their hands, even though they are wearing an audio headset. For this reason, the authors advise this kind of safety gesture to trigger safety features and presented a "Close your ears" gesture that allows the user to mute everyone nearby by keeping the controllers next to their ears for three seconds.

3 Taxonomy of the Safety tools

First, as mentioned earlier, we define a Safety Tool as an Interaction Technique that aims to modulate social interactions between the users, in order to mitigate situations that they perceive as inappropriate.

In this section, we present a taxonomy of Safety Tools. The process of making the taxonomy was inspired by the work of Bowman et al. [12] who attempted to formalize the design of 3D interaction techniques. The goal of the taxonomy of safety tools is to decompose the safety tool into several basic subtasks (each of which represents a decision that needs to be made during the design of the safety tool), and list several possible technical components for each of the subtasks, and also some dimensions that characterise these subtasks. These lists do not claim to be exhaustive.

3.1 Method

The creation of the taxonomy is based on the method presented by Nickerson et al. [36].

First, we identified the meta dimension of the taxonomy, which is its main goal: to provide a decomposition of the safety methods to researchers and designers of social virtual reality applications. In the second step, we conducted an Empirical-to-Conceptual Iteration in order to provide an overview of the subject, in which we identified the safety methods present in the three main SVR platforms: VRChat [5], Rec Room [3] and Meta Horizon Worlds [1]. This included (1) testing the different safety features in the three SVR platforms, (2) reviewing the platforms' website pages about safety [2, 4, 6], (3) reviewing the literature on harassment in Social VR, and (4) strategies used to mitigate it. This process provided a first overview of the main dimensions of the safety techniques considered in the taxonomy. From this first iteration, we found that the activation process of a safety tool can be decomposed into (1) the activation of the tool, (2) the selection of the target, and (3) the manipulation/modulation applied to the social interaction.

The third step was to understand the concept of the interaction technique, in order to use it to describe the safety methods. To achieve this, we conducted a Conceptual-to-Empirical iteration, based on the Task analysis method and the Taxonomy of Interaction techniques presented by Bowman et al. [12]. Based on the third step above, we thought it necessary to decompose the task accomplished by Safety Features into three different subtasks: (1) the selection of the target, (2) the selection of the manipulation/modulation, and then (3) the selection

of the parameters of this manipulation and their application. We used the classification of selection from their work [12] to describe the selection of the target and the one of the manipulation/modulation for safety tools that are manually activated by the user. Therefore, we wanted to better describe how safety tools can be used automatically as proactive measures, and what kind of manipulation can be applied on a social interaction (and not only check what is currently available in the platforms). Therefore, the selection process of some safety tools did not fit into the classification of Bowman et al., because they are mainly automatic, independent of the user decision (like the Personal Space Bubble of VRChat, which selects all users in a certain radius from the user). Besides, some social interaction manipulation can be described as the manipulation of a 3D object (changing the position of the target, in the case of the Push-away gesture [53]), but this is not possible for the majority of manipulations. Thus, Zheng et al. [53] used the notion of *social cues* to analyse inappropriate behaviours in social VR and map how currently available safety tools can be used as mitigation. We followed their steps and tried to describe the workflow of an automatic safety tool as selecting a target based on social cues they are emitting, selecting the appropriate manipulation based on the same cues, and then applying the manipulation, which can be considered as a modulation of the emission or perception of these social cues. To achieve this, we conducted a Conceptual-To-Empirical iteration, based on articles about social interactions, social cues, and computer-mediated communication [17, 26, 38]. With this iteration, we have identified that the selection of the target and the selection of the manipulation/modulation can be automatic by detecting two types of factors: social cues coming from the target and arbitrary tags assigned to the target.

Some social cues are linked to inappropriate behaviours, as shown by Zheng et al. [53], and some safety tools are detecting them to act proactively (like the Personal Space Bubble of VRChat [5], selecting users which are too close: a proxemic cue. Also, others are using a tag system, letting users assign tags to others and then automatically select targets wearing these tags and select an appropriate manipulation. For example, the Trust and Safety system of VRChat [6, 13] automatically assigns a rank to users, depending on how they behave, and lets them decide to filter out some interactions coming from users, depending on their rank. Also, in Rec Room, users can tag others as their friends to only receive private text messages from them (adjustable in the *direct message preferences*) [4]. This is because the context of a social interaction impacts its perception, and the context (e.g., relationship between participants) can be expressed easily with tags (e.g., the tag 'friend'). Lastly, the manipulation applied by a safety tool can be described as a modulation of the emission or perception of a social cue, which is avatar-mediated through the Virtual Reality possibilities: e.g., the Mute system of VRChat is a modulation of the user perception of auditory social cues (volume turned to 0) coming through the voice communication channel of the target. This last iteration, which resulted in separating the concepts of automatic selection (detection), and what is impacted by the manipulation (the emission or perception of social cues), led us to develop a new characterisation of social interactions in SVR, which will be presented in the following section.

3.2 Characterisation of inappropriate social interactions

While there exist previous characterisations of inappropriate social interactions in SVR, they are either based on the harmful content of the interaction (e.g., violence, scarring, sexual harassment) [53], on the modalities used to convey the interaction (verbally, physically, or through the environment) [11] or on the sensory modality (verbal, meaning auditory; visual and physical) [51]. These classifications focus on which types of social interactions are perceived as inappropriate, but do not allow for a precise description of how the aggression is performed by the harasser neither how it is perceived by the victim.

To understand how we can model inappropriate social interactions in SVR, we will first focus on how social interactions are modelled in real-world interactions. We will base our model on the linguistic model of verbal communication of Jakobson [26] which we will extend with works from social psychology [38], to take into account nonverbal communication and to add sociological and psychological factors that have an impact in the context of the interaction, like the relationship between participants. The communication process can be modelled on the transmission of a *Message*, being emitted, intentionally or not, by a *Sender* to another user, the *Receiver*. In case of verbal communication, the meaning of the message is "encrypted" by the use of a *Code*, and the message is emitted in a *Context*, by the means of a *Contact* between the protagonists. Finally, the message is decoded and infered with the Code and Context of the receiver, to extract a meaning. In the following, we will discuss the importance of the *Code*, the *Context* and *Contact* in social interactions in SVR.

3.2.1 Codes shared between participants

In linguistics, the *Code* is a conventional set of signs that are combined to convey a meaning. Codes in SVR are common to reality, as they are mainly linked to language elements: oral, written, and gesture-based. Because of the international access to Social VR platforms, users coming from all around the world can interact with each other. This can lead to situations where an interaction could be perceived as inappropriate because of a mismatch in the *Codes* of the participants. Participants may not speak the same language, or may not have the same speaking skills of a shared language. Thus, an inappropriate situation could emerge unintentionally, due to difficulties in understanding each other.

3.2.2 Context of the social interaction

Picard [38] described the *Context* of a social interaction as the sum of all elements that are giving meanings to the interaction, but are not part of it. Originally described by Jakobson [26], the context is composed of linguistic elements that serve as a referent of the communication; all mentions of factual information of the outside world. However, linguistic elements are not enough to describe social interaction, since the social condition of the environment also has an impact, Picard presents the *Situation* as one part of this context.

The Situation is composed of the (a) Frame, the physical, topological and temporal aspect of the interaction; (b) the Institution, the norms and constraints inherent in the cultures and social group that are involved; (c) the Participants, their number and role in the interaction; and (d) the Relationship between the participants, formal or informal. All these elements can have an impact on the perception of a social interaction as being inappropriate. For example, an insult coming from a friend can be perceived as a joke, but the same coming from a stranger is more likely to be perceived as aggressive/confrontational.

The notion of Context is an important aspect of social interactions in SVR platforms. Currently available safety features in SVR platforms and proposed ones address the different elements of the Context. Regarding the Relationship dimension, roles/tags among users can be used to determine the available interactions among users. For example, VRChat uses the *Trust and Safety System* to categorise users among different roles: Visitor, New User, User, Known User, Trusted User and Friend. This tag system defines the Relationship among users and determines the available interactions, which can also be customized. Moreover, SVR platforms inform users regarding existing rules and norms, by making them sign a Code of Conduct, to make sure that all users are aware of the Institution. Also, some behaviours that became common in SVR could be considered as inappropriate in Reality and for newcomers, such as *head patting* someone you just met in VRChat. Regarding the Frame, Schulenberg et al. [43] suggested customising the activation of safety tools based on the type of virtual environment: the protection should be more restrictive if the user is in a public world or the private world of a stranger, and less in his private world or one of their friends.

3.2.3 Contact between participants

The *Contact* in a social interaction, is the physical channel linking the sender and the receiver, used to convey the message. Social interactions in VR are mediated by the technological device, but at the same time, they aim at reproducing real face-to-face social interactions. Thus, the VR technology used greatly impacts/modulates the Contact between

users. The Contact can further on be divided into two actors: the *Sender*, which uses the interaction capabilities of the Avatar-mediated communication in SVR to initiate the contact, and the *Receiver*, which perceives the contact through their different sensory channels, as a social cue. We first explore what the sender can manipulate in order to generate an inappropriate social interaction to the receiver. Based on this, we can design appropriate safety tools with the goal of mitigating such behavior.

Sender Communication Channels From our observations of the currently available SVR platforms, and from our literature review, we build a simple model of the different means of the Avatar-Mediated Communication accessible to a sender in SVR.

Figure 1, summarises the different means of communication accessible to a sender, to emit, intentionally or not, a social cue that could be perceived by a receiver. The first mean of communication in SVR is the user's avatar, "a perceptible digital representation whose behaviors reflect those executed by a specific human being" (Bailenson and Blascovich 2004 [10]). This representation can be separated in three parts: (1) the Computer-Mediated Communication Channels (or CMC Channels), that can be textual or oral, between users, (2) the 3D Representation, such as the user's Avatar, and (3) the Movements of the avatar in the virtual environment. We base our idea of this 3D representation on what is currently available in SVR platforms. In *VRChat*, the creator can add components to it: audio sources to play predefined sounds, blasters of particles, and interactable zones to trigger visual, haptic, audio, or even olfactive feedback. These three parts of the representation can convey inappropriate interactions: a User can injure another by using the oral communication channels, or by doing offensive gestures with their avatar, or by embodying an inappropriate avatar, with insults embedded in its texture.

The second means of communication is through the **Interactable Contents** in the virtual environment, which are shared with others. In fact, these contents can serve as a weapon to harm others: a user can throw objects at their victim [18], try to force someone to eat a stick [37] or destroy what they are doing, like a pile of blocks [11, 45]. In this category, we can further distinguish **User Generated Content** (UGCs) from the rest of the content, because of the link between this content and the user, but also because of the freedom of expression allowed by the free creation of the content. Indeed, a user in SVR can grab a virtual pen, an already existing interactable content, and start drawing inappropriate content in the shared virtual environment [37].

This distinction between direct interaction coming from the avatar of the sender, and indirect coming from their interacting with the environment is important because, even if all means of direct interaction are blocked, the sender is still capable of interacting with the receiver by using the environment [14].

The last means of communication is through interacting indirectly through the **real environment**. This can be by exploiting system vulnerabilities or known issues. For example, the "crashers" attack in VRChat [21, 41, 51], which makes the HMD of the victim crash by displaying with their avatars, textures impossible to render. Another example is through Virtual Physical Perceptual Manipulation (VPPM) [46], like using redirect walking to make the user run into a real wall. Because this part is more linked to the security of devices or applications and cannot really be moderated with safety interaction techniques, we will not further consider such interactions in the remainder of the paper, yet additional studies are required to fully understand how to face these risks.

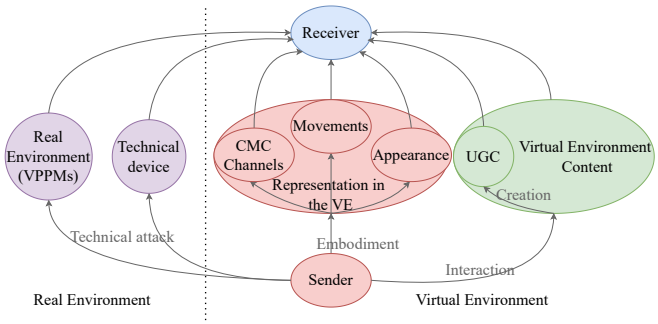

Fig. 1: Diagram summarising the communication channels available to the Sender to socially interact with the Receiver in an SVR context.

Receiver Sensory Channels The receiver perceives the message(s) of the sender with their senses and interprets them as *social cues*. Feine et al. [17] proposed a classification of the social cues involved in computer-based communication with a communicative agent, which can be translated with the virtual avatar of a user or an NPC. The classification considered four main categories: verbal, auditory, visual and invisible, but, in order to emphasise the importance of channels of communication, which can be modulated by safety tools, we propose to classify social cues based explicitly on the sensory modality (see Figure 2). First, we classify social cues based on the four main senses that are linked to VR experiences: (1) the sight, i.e., visual cues; (2) the Hearing, i.e., auditory cues; (3) the touch, i.e., haptic cues; and (4) the smell, i.e., olfactory cues. Furthermore, to account for the invisible cues from the taxonomy of Fiene et al. [17], and to ensure consistency, we included it, but renamed it as (5) non-sensory. This category includes social cues that are not perceived directly by one of the senses, but unconsciously, like feeling that someone is present, knowing that someone is in the virtual environment without any cues, but can trigger social reactions [14]). Another non-sensory cues are chronomic cues, the sense of passing time, for example, receiving too many messages in a short time: *Spam*.

Then, we propose to further split them into two categories: verbal cues that are coded with a language, and non-verbal cues that do not involve a language. These verbal cues can be visual cues, with gesture-based or writing-based languages; auditory cues, with speaking-based languages; or haptic cues, with touch-based language (e.g., Braille). As presented by Feine et al. [17], a verbal cue is composed of the *Content*, the strict and literal meaning of the message and the *Style*, the meaningful deployment of language variation in a message. For example, the perception of a verbal interaction as inappropriate could be because of insults, the Content, or if the message is expressed sarcastically, the Style.

For non-verbal cues, we further subdivide the different sensory cues: Visual cues can be divided into (a) Kinesic cues, linked to body movements, like offensive hand gestures or unwanted touch, (b) Proxemic cues, based on personal distances, physical positions of users in the virtual environment, like personal space invasion; (c) Appearance cues, based on the appearance of the avatar or virtual agent, like the use of a sexually offensive avatar; and (d) computer-based communication cues, visual elements that can enhance or modify the meaning of a verbal message, like sexually explicit emoticons.

Then, auditory cues are sorted into (a) sound qualities, permanent and adjustable characteristics of speech or sound, such as a voice volume which is too high, and (b) vocalisations, non-linguistic vocal sounds or noises, such as *Wolf Whistling* someone.

Haptic and olfactory cues are strongly linked with the hardware being available for the Receiver. This contrasts with visual and auditory cues that are always available in SVR platforms. Here, we provide a first classification taking into account key dimensions of haptic and olfactory devices. Regarding haptic cues, we can consider two main dimensions, (a) the haptic modality and (2) the localisation. The haptic modality is linked to the organs that are being stimulated by the haptic device, from kinesthetic feedback (muscles, tendons) to tactile feedback (mechanoreceptors in the skin). The types of sensations that can be rendered can range from force or vibrations to thermal or tingling sensations. The localisation dimension defines the location over the user's body in which the sensation is rendered, which can range from single-point stimulation (vibrator in a VR controller) to larger parts of the user's body (vibrotactile vest). Regarding olfactory cues, we will limit the classification to the type of scent.

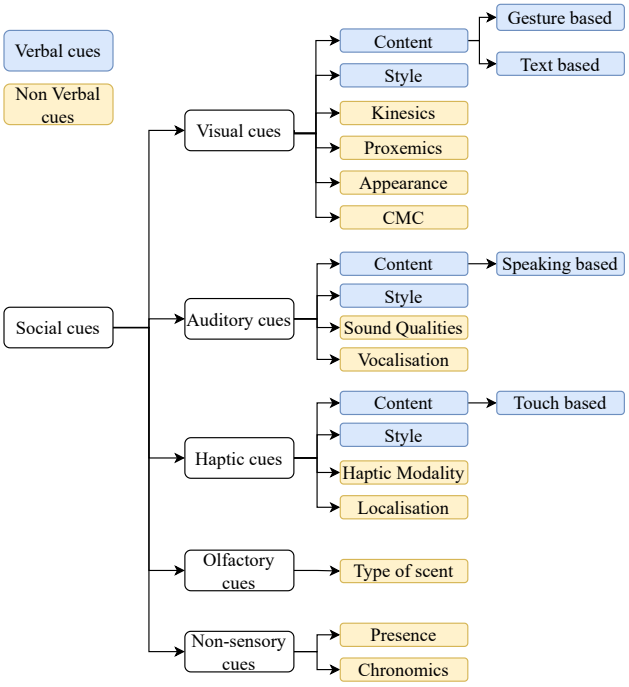

Fig. 2: Classification of the social Cues that can perceived by the Receiver of an interaction in SVR focusing on the involved sensory cues.

3.2.4 Illustrative example

This section provides an example of how the taxonomy can be used to analyse a number of social interactions that could result in a situation of sexual harassment.

Let us take a concrete example of sexual harassment in SVR. The user is confronted with another user in a form of an avatar who is sending inappropriate jokes with sexual innuendos as text-messages. He makes several jokes, one after the other. The avatar also shows emojis with sexual connotation in the text-communication channels (e.g., eggplant emoji). This interaction combines Visual Verbal cues (making jokes), Chronomics (repeating jokes in a short period of time) and Visual Non-verbal cues (emojis). This example also shows the inappropriate communication involving the CMC channel.

Now let us consider a harasser who is whistling after the victim, is showing inappropriate hand gestures and in the end, grabs a wooden stick which is in the environment and touches the victim in an inappropriate place on the avatar's body. Inappropriate gestures are Visual Non-verbal cues, induced by the movements of the avatar, where whistling after a victim is an Auditory cue, more specifically Vocalisation, passing through the communication channel for voice. The loudness of the voice is a manipulation of Sound Qualities in the Auditory cue category. The unwanted touch by the use of a Virtual Environment Content (the wooden stick) is a combination of the perception of a Visual cue (seeing the wooden stick coming in contact with the avatar, so Kinesics) and, if the victim is wearing a haptic vest, the perception of a vibration in an inappropriate Localisation. In addition, the harasser used the movement of their avatar to invade the personal space of the victim which is the example of Proxemics in the Visual cue category.

3.3 Task decomposition of safety tools

In the previous section, we detailed the communication process and the different social cues that are involved in it. The objective of a safety tool is to manipulate the social cues that are conveyed by a means of the Avatar-mediated communication, either by the Sender or the Receiver, in order to mitigate/avoid inappropriate social behaviours. Thus, the safety tool should allow users to precisely identify *who* generated the inappropriate social behaviour, *what* social cue should be manipulated and *how* it should be modulated. In the following, we will employ the term *User* to name the person using the safety tool, and the term *Target(s)* refers to the entity (user, group of users, or virtual content) that is being affected by the safety tool.

Taking inspiration from the work of Bowman et al. [12], we propose a task decomposition of a safety tool technique into six different subtasks: (1) the selection of the Target, (2) the selection of the Contact of the interaction, (3) the selection of the parameters to modulate, (4) the selection of the strength, (5) the selection of the direction, and (6) the selection of the duration.

As an illustration, let's consider that a user wants to mute another user in VRChat. (1) First, the user has to access the target's profile page. The profile page can be accessed either by selecting the user in the list of nearby users through a 3D menu, or by directly pointing to the user using raycasting selection. (2) In order to select and trigger the manipulation of the auditory cues coming from the voice communication, the user has to push the mute button in the user's profile page. (3) The manipulation is to set to zero the volume of the sound, (4) that is perceived by the User, (5) emitted by the Target, (6) until a revocation from the user. In the following, we further detail the different components for each subtask.

3.3.1 Selection of the Target

Figure 3 summarises the different components of the taxonomy for the Target selection. When selecting the target, two main alternatives can be considered: (1) the selection is explicit, the user manually selects the Target, or (2) the selection is implicit, automatically defined given a series of rules and heuristics.

The *Manual* selection of a target is a classical selection of a 3D object: the 3D model of the avatar or the virtual agent. Typical selection methods in SVR consider indirect selection through the use of GUI, a notification triggered by a bystander intervention [28], or direct selection, such as pointing the Target with a ray. However, any 3D object selection method could be used for such purpose [9]. Yet, manual selection methods are typically reactive, as they are initiated after the inappropriate social interaction has already been initiated.

In order to anticipate the inappropriate behaviours or rapidly react upon them, *Automatic* selection methods are normally considered, in which the targets being selected are defined by the creation of rules linked with Social cues. For example, a rule based on Proxemic cues can be used to select all users that are at a certain proximity of the user; this selection method is used by the *Ignore Bubble* Rec Room [3]. Other rules can also be defined to select users swearing (Auditory cue, Content), performing offensive gestures (Visual cue, Kinesics) or spamming textual messages (Invisible cue, Chronomics).

Automatic selection methods can be customised at different levels, either by the user (User-defined) or by the SVR platform (System-defined). For example, the radius of the Ignore Bubble of RecRoom is User-defined, while in VRChat the radius of the bubble is imposed by the system. System-defined configuration provides a default system behaviour while still allowing users to further configure the system based on their preferences.

Finally, some of these rules can be further configured considering the Relationship dimension. The most prominent example is the *Tag System* in which each user can be tagged in order to determine the role of the user with respect to other users or within the system. For example, with the Trust and Safety System of VRChat [5, 13], users can assign the tag *Friend* to others, and the moderation system of the platform determines the *Trust Ranks*. The trust rank is supposed to represent the trustworthiness of a user, by being an indicator of their commitment to the platform and to others. This rank starts with "Visitor" and then falls to "Nuisance" and climbs to "Trusted Users". If we revisit the example of the personal bubble, users could define different radii depending on the tag of the other users: a small radius for "Friends" and a larger radius for "Visitors". Moreover, based on the Relationship dimension, the user could also define to manually select all users belonging to a specific group, or even select all users. For example, the Safe Zone of Meta Horizon Worlds [1], which selects all users in the virtual world after the press of a button.

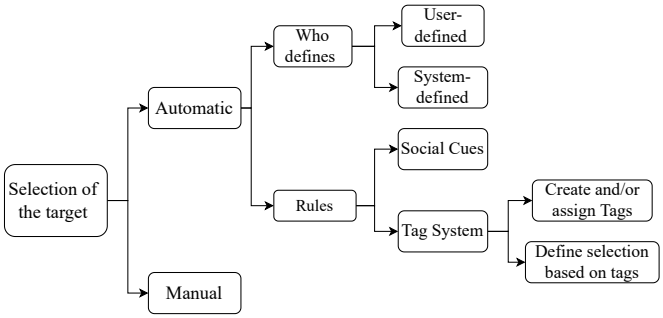

Fig. 3: Taxonomy for the Target Selection, inspired by the decomposition method of Bowman et al. [12]. The manual modality is not developed as any object selection technique method can be considered [9].

3.3.2 Selection of the Contact

The next step is the selection of the *Contact* that will be manipulated. As mentioned before, the contact is a combination of a possibility of the SVR Avatar-mediated communication, which is used to create the interaction, and of a social cue, which is what is perceived from the interaction. For example, the Mute system of VRChat selects the Auditory cues coming from the computer-mediated communication channel for voice.

When selecting the contact that should be manipulated, similar to the selection of the Target, two main alternatives can be considered: (1) the selection is explicit, and the user manually selects the manipulation to be applied, or (2) the selection is implicit, and it is automatically selected given a series of rules and heuristics. However, compared to the selection of the target, the Social cues that can be manipulated are limited and well defined.

The *Manual* selection of the Contact typically requires the use of a GUI, and selecting the Contact to be manipulated. For example, selecting the Mute button in the profile menu of the Target (*RecRoom*). However, other methods can combine the selection of the target and of the contact in a single action. The *Talk-To-Hand* in *RecRoom* is activated by orienting the palm of the hand towards a user, with the fingers pointed to the top. In this case, the hand orientation is used to select the target, and the hand gesture defines the contact to be manipulated.

Regarding *Automatic* methods, they can also be defined based on a set of rules that can be User-defined or System-defined, and further modulated by Relationship dimension. Methods that automatically select the Contact are usually linked with an automatic target selection method. For example, let's consider the following rule: Make transparent all the non-friend users who are closer than 1m. This rule defines both the targets and the manipulation to be applied. Another example is the *Trust and Safety System* of *VRChat* which uses an automatic selection of all users, and then automatically selects the Contact to manipulate. Depending on a configuration defined by the user: they can, for example, manipulate the Auditory cues coming from the computer-mediated communication channels, or from the Appearance (with audiosources); or Visual cues coming from the Appearance of the avatar of the target.

3.3.3 Selection of the modalities of the manipulation

After selecting the Contact, the *modalities* of the manipulation need to be selected. For each of the social cues linked to a sensory modality, some manipulations are possible. The content of verbal cues can be replaced by another one (changing insults with the asterisk symbol), and sound qualities could have their volume or their pitch changed (a mute of auditory cues is a volume turn to zero). The level of transparency or the size of the Visual Appearance of an avatar can be modified, and the intensity or the localisation on the body of a Haptic feedback could be modulated (to move it to a less sensitive body part, for example).

The selection of the modality, in most of the cases, is done in a UI by the user, or by the designer of the safety tool, and imposed on the user. The modalities can be selected in three different ways, depending on the data type of the parameter. It can be only a switch between ON/OFF (the *mute* of RecRoom, is either turning the volume to zero or to normal) or a choice between an enumerable list of options (the *Voice Canal* of Meta Horizon World lets the user choose if they want to turn to zero, for them, the voice volume of non-friend users, or to replace the content of the verbal communication with gibberish, or no manipulation at all). Finally, this selection can be done by choosing an 'analogue' value in an interval: using a slider to select the perceived volume of the voice of others.

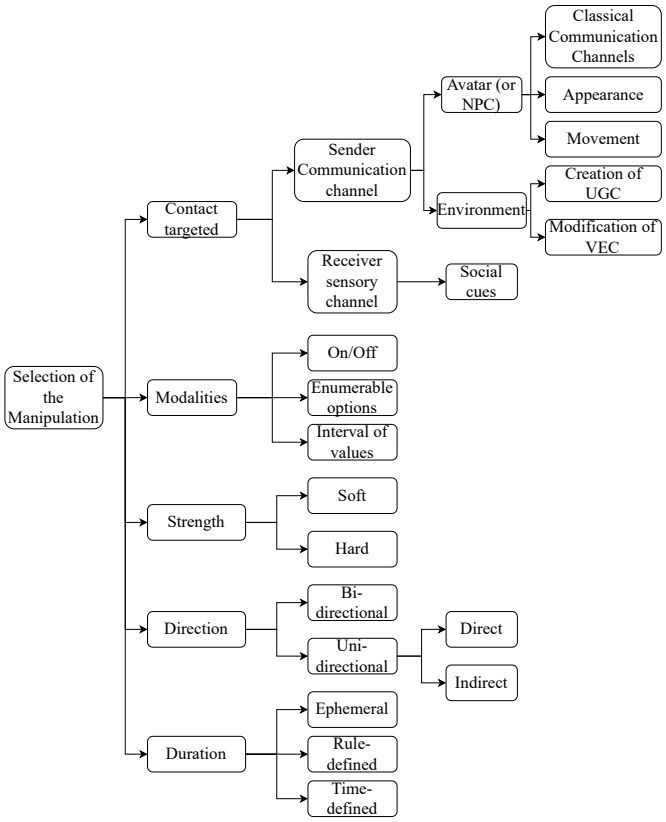

Fig. 4: Taxonomy of the manipulation selection inspired by the task decomposition method of Bowman et al. [12].

3.3.4 Strength of the manipulation

Manipulation can be divided into two groups, depending on which part of the interaction is manipulated. *Hard* manipulations are impacting the capacity to interact of the target, like the automatic *Voice Moderation* from Rec Room [3], which can impose a volume turned off for the oral communication ability of the target. These manipulations are mainly used to punish the target and protect other users, so their use should be limited to members of moderation systems or to a community use (like a *Votekick*) [16]), to avoid risks of abuse and new forms of harassment.

In contrast, *Soft* manipulations impact the user's (or the target's) perception of the interaction. These manipulations are the most common in currently available platforms, because they are not restrictive for the target. The risk of misuse for this kind of manipulation is low, but still exists (like the social exclusion of someone in a group). These manipulations are interesting for tools that can be used by every user, to let them express their preferences without impacting others. For example, in Rec Room, a user can *mute* another, and thus turn off the volume of oral communication that they will perceive from the target. The selection of the strength is usually done by the designer of the safety tool, and imposed on the user.

3.3.5 Direction of the manipulation

Due to the possibilities of VR and, more generally, of Computer-Mediated Communication, the physical perception of a social interaction of each actor can be modified to be different from the one of the others. We can identify two main directions for the manipulation: the manipulation can be *Unidirectional*, so applied on interactions coming from one actor to the other, or *Bidirectional*, so applied on interactions coming from both actors.

We can identify two types of Unidirectional manipulations: direct and indirect. Unidirectional Direct methods only affect the User by masking what is considered by them inappropriate behaviour coming from the Target. This kind of manipulation can be observed in the mute system of VRChat, which mutes the voice channel of the Sender. In contrast, Unidirectional Indirect methods are aimed at the perception capacity of the Target, for interactions coming from the User. Their objective is to protect the user by hiding them and/or what they are doing from the perception of the target. This type of manipulation can be found in systems that allowed users to mute themselves, or quick travel systems [7] that are used to fly away from an inappropriate social interaction, by going to another virtual environment.

Finally, manipulations can be Bidirectional, applied to the perception capacity of both the user and the target for interactions. The goal here is to protect the user by both the means presented above, hiding the inappropriate behaviour from the user, and hiding the user from the target. An example of these manipulations is the *Block* of VRChat, which makes the user and the target both invisible to each other. It is important to note that Bidirectional manipulations are considered better for the protection of users, especially for block systems. In fact, as shown by Chen et al. [14] even if a user has blocked their harasser (and made them invisible), they can still perceive their presence by indirect interaction (through the virtual environment or other users). Thus, hiding themselves from their harasser enforces their protection by making them harder to localise in the virtual environment. The selection of the direction is usually done by the designer of the safety tool, and imposed on the user.

3.3.6 Duration of the intervention

The last dimension of manipulations in our taxonomy is the duration, which defines the time in which the method is active. We consider three different possibilities: ephemeral, time-defined, and rule-defined. For *Ephemeral* manipulations, they only last for the duration of the inappropriate social interaction, like the *Push Away Action* [53], or are instantaneous, like using the *Quick Travel* to fly away from the situation. *Time-defined* manipulations are active for a fixed period, as the *Votekick* in VRChat, which lasts for one hour. This time period can also be undefined and active until the User revokes the manipulation, like the *Block* in VRChat, which lasts until the user pushes the unblock button in the profile of the target. Finally, *Rule-based* manipulations are active while a predefined rule is valid, like the *Personal Space Bubble* from VRChat, which makes the avatar of the target invisible as long as they stay near the user. The selection of the duration is usually done by the designer of the safety tool, and imposed on the user.

3.4 Ilustrative example

We classified the safety tools we identified into our taxonomy. For example, the characterisation of the safety tool *Mute*, available in the platform VRChat, is:

The selection of the target is manual, either in a list of all nearby users or by pointing with a raycast at their avatar. The selection of the contact (which is the Audio cues coming from the voice communication channel) manipulated is done manually, by clicking a button in the profile page of the target. Then, the selection of the different characteristics of the manipulation (parameter, strength, direction, and duration) is imposed on the user by the platform: the manipulation is turning the volume to zero, and is Soft, Unidirectional Direct, and Time-defined (Permanent).

The full table is available in supplementary materials.

4 Discussion

The main contributions are threefold. First, we made a characterisation of social interactions in Social Virtual Reality, based on models taken from social psychology, which were adapted to the particularities of computer-mediated social interaction in VR. Second, we used this new characterisation to propose a taxonomy of safety features, which are interaction techniques for manipulating social interaction between users, and a way to mitigate a situation perceived as inappropriate. Lastly, we classified existing safety features into our taxonomy of safety techniques.

The advantage of the classifications we provide is that they are not based exclusively on user studies, but rather on theories that have been developed in the past by different disciplines: communication and interaction models from social psychology and linguistic sciences [26, 38], the interaction techniques model [12], and the taxonomy of social cues for conversational agents [17]. This enables us to expand on the terminology of safety tools without introducing new terms.

Furthermore, our taxonomy allows a precise decomposition of the *activation process*, the Selection of the Target and the Selection of the Manipulation of safety features, which was not done in the past. The classifications of Zheng et al. [53] and Weerasinghe et al. [51] are not making the distinction between features that are similar in their concept but different in their activation process, such as different ways to mute someone in the *Mute and Block* feature [51] or the ways to trigger the *Safety Zone* [53]. For example, in Rec Room, a user can mute another by pressing a button in their profile, but he can also open the profile menu by pointing to their avatar or by navigating in a user list. He uses the same Manipulation: a mute of verbal communication coming from the target, but with a different activation process. This is useful for evaluating the performances of a feature, because the activation process can play a significant role, especially when the inappropriate social behaviour is causing significant stress to the user, as shown by Zheng et al. [53]. In those cases, users will react automatically, using natural defensive gestures; thus, the suggestion of the authors to use natural gestures as triggers should be considered as an activation process.

In contrast to the two previous classifications [51, 53], our taxonomy also allows a more precise characterisation of the Manipulation that is applied on the interactions in SVR, such as the example of blocking another and the avatar appearing as a grey and semi-transparent silhouette (RecRoom), as opposed to becoming fully invisible in VRChat. Also, the manipulation in this example is Unidirectional-Direct in RecRoom, but Bidirectional in VRChat. Even though these two features are called "Block", the manipulation from VRChat offers more protection than the one in Rec Room.

Thus, our taxonomy is important for the evaluation of the performances of safety features, because it allows the identification of various factors that could have an impact on their use, such as how the target is selected. To our knowledge, the first example of such a study on safety tools is the work of Fiani et al. [18], in which they compare the perception of children that were using three different activation processes for *blocking* a group of harassers: selection from a list of users, selection of all users, and selection by pointing at the user's avatar. They demonstrated that children prefer and feel safer when they make a selection by pointing at the harasser's avatar, rather than selecting him from a list of users.

Our taxonomy also emphasises the role of the automatic selection of the target and the manipulation, and the role of the pre-selection. Many studies focus on automatic detection, using AI, for inappropriate behaviours in text [32, 35] or images [47], but only a few of them are trying to detect inappropriate behaviours caused by the interaction possibilities of the SVR medium. To our knowledge, the only example of such studies is by Wang et al. [50], in which they identified inappropriate social interactions in Social Virtual Reality using the relative avatar position of the users, *Proxemics*, and the position of their fingers given by actions on controller buttons and hands (*Kinesics*).

4.1 Identifying issues of existing safety tools

In this section, we discuss a number of issues that can be raised by the use of safety tools in the context of SVR platforms, which were

identified during the creation of the taxonomy.

Manipulation of the virtual environment by blocked users: Even when a *blocked* user is fully invisible, the actions of a target can still influence the content of the virtual environment. Existing platforms should consider putting the restriction also on UGC or limit the use/modification of the virtual environment content. Quick travel or ban are available, but they force total disconnection, which is a severe manipulation, which could prevent the users from feeling secure in the platform.

Pre-defined Restriction Rules: Even if several automatic safety systems allow users to pre-define the selection of the target, only a few let them customise the manipulation and prefer to impose a strict restriction, such as a full mute of the verbal communication or making the appearance of the avatar fully transparent. Furthermore, most of the manipulations of the safety features are *Permanent* or *Rule-Defined*, none of them allow customising a fixed time period for the application of the manipulation. Most platforms rely on manipulation that does not allow enough nuances in the restriction of the interaction, as they are too restrictive [14, 43]. The design of safety tools with a more precise Manipulation strategy could lead to a better balance between the benefits of the social interaction in SVR and the protection of the user from inappropriate behaviours of others.

Asymmetric content: The use of unidirectional manipulations can cause an issue of a virtual world that is asymmetric for all users, and these could lead to some issues, depending on the manipulation. As presented in the issue of the block system, a *Soft Unidirectional-Direct* manipulation that forbids the *target* to take in their hands a virtual environment content will create a problem when the target takes a virtual object: the target will have the block in their hand, but the *user* will see the block still on the ground. Then, two instances of the same object coexist in the virtual environment, but one is only accessible to the user, and the other to the target and all others. These lead to issues of asymmetric VR experiences.

4.2 Guidelines for the design of safety tools

In this section we discuss a number of guidelines that were extracted from the presented taxonomy and related to related works.

Freedom of customisation: Following the guidelines of Reid et al. [39], the users need to "feel good and in control" to have a good experience in SVR. If not, they tend to leave the platforms to not face toxic behaviours [7, 31]. At the same time, previous studies [28, 43] highlighted that users need to be free to customise their safety features; if not, their experience will be restricted by the system's predefined parameters, and they may choose to turn it off completely (like the VRChat safety bubble, which will turn all nearby avatars invisible in a certain radius). It is also beneficial to promote the automatic tools for which parameters for the Selection can be user-defined rather than system-defined. For example, users do not always want permanent disconnection from others and prefer a temporary solution, which can send a warning that this interaction has been perceived as toxic [14]. Thus, designing safety features with manipulations where the duration is ephemeral could be a solution. For example, by automatically selecting users, based on a detection of the Content of Auditory Verbal cues, could detect insults, and with a manipulation of the sound qualities of the Vocal communication it could mute specifically the insults. With the continuous choice, the manipulation could last for a fixed period, predefined by the user, for example, a mute/block for one hour.

Investigating new selection processes: The taxonomy shows the importance of the selection processes in the use of safety tools. We presented the manual selections and the automatic selection by detecting social cues coming from the target, or tags that have been assigned. This list of selection methods is not exhaustive, and analysing how human users react in a situation of stress caused by an inappropriate interaction may be a good solution for developing new types of selection processes. Gesture-based activation was used by Zheng et al. [53], highlighting how the use of natural defensive gestures, like covering the ears as a protection from loud music, can be used as triggers for safety tools.

Adapting the selection process to the harm: Another natural reaction to facing a toxic situation, observed in the literature in psychology, that has a huge impact on how a user can defend themselves, is "tonic immobility" [48]. This reaction consists of a total freeze of the body of the victim, and is more likely to happen during sexual and traumatic interaction. In this situations, using a quick reaction tool that requires a movement or a press on a controller to be activated might not be adapted. Methods capable of detecting a freeze of the movement of the user (a reduction of body sway) associated with physiological measures of stress (a decrease of heart rate variation and an increase of heart rate, if the user is wearing a smartwatch, for example) and eye movement (if the headset is equipped with an eye tracking device) could be studied.

Privacy and risks of abusive regulations: In SVR platforms a large amount of user data can be collected. With a precise detection of social cues, it is possible to collect and analyse private information of the users, such as their age or gender. This information then can be used to apply manipulation on interaction with these users, and create new forms of discrimination, where users from different communities share the same virtual environment, but can't interact with each other [13, 14]. It is important to mention that the creators of the platforms have their own private interests, which affect the functioning of the SVR system, and they define what kind of behaviour is acceptable or not. This can have a huge impact on which safety tools they create. Further research on the impacts of social behaviour is therefore important to inform the owners of the platforms about the possible impact their system could have on society.

5 Limitations and Future Works

This study addresses a relatively recent field, where empirical research still remains limited. Our analysis primarily focuses on the most adopted SVR platforms, selected for their relevance and real-world use. This scope also motivated the development of the proposed taxonomy, which aims to structure the understanding of safety risks in these environments. However, the study's focus on relatively mature platforms, characterized by large user bases and iterative safety improvements, constrains its generalisability. Future research should extend this framework to lesser-known or emerging SVR platforms, offering a more comprehensive analysis of safety challenges across diverse user populations and technological contexts.

While our taxonomy of avatar-mediated communication possibilities in SVR is grounded in currently available features, it can also anticipate emerging modalities, such as haptic or olfactory interactions, as well as real-world interactions like perceptual manipulations or system vulnerabilities. This approach ensures the taxonomy's adaptability to technological advancements. Nevertheless, the rapid evolution of SVR platforms may introduce novel communication channels or manipulations that fall outside our current model. Ongoing research is essential to identify and integrate these possibilities, ensuring the taxonomy remains robust and relevant as the field progresses.

The scope of this work is limited to Social Virtual Reality, yet its conceptual framework could be extended to augmented or mixed realities (AR/MR), in which participants could share or not the same physical space. When co-located, however, new manipulation strategies, such as the concealment of real-world objects, may emerge, necessitating further theoretical and empirical exploration. Additionally, certain manipulations, like the alteration of haptic cues, may prove infeasible in shared physical spaces, as users retain direct tactile access to one another. Addressing these context-specific challenges will be critical for adapting the framework to AR/MR scenarios.

Finally, at present, our model remains theoretical, and its practical relevance, particularly the significance of its dimensions, has yet to be empirically validated. Future studies should prioritize formal evaluations to assess the taxonomy's applicability in real-world SVR applications. Furthermore, while this paper does not examine the occurrence of safety risks, identifying which communication possibilities and social cues are most strongly associated with inappropriate interactions is a vital next step. Such research could inform the development of targeted mitigation strategies and contribute to the creation of a standardized evaluation testbed.

6 Conclusion

In this paper, we presented a new taxonomy of safety tools specifically for Social VR. While this is not an exhaustive classification system, it provides a systematic step to organise the existing terminology in the area of safety tools and provide future directions.

Acknowledgments

This work was supported by the project META-TOO, funded by the European Union's Horizon Europe research and innovation program under grant agreement No. 101160266.

References

[1] Meta horizon homepage. https://horizon.meta.com/. Accessed: 2025-08-26. 1, 3, 6

[2] Meta horizon safety page. https://www.meta.com/fr-fr/help/quest/1737463343292580/. Accessed: 2025-08-26. 3

[3] Recroom homepage. https://recroom.com/. Accessed: 2025-08-26. 1, 2, 3, 6, 7

[4] Recroom safety page. https://recroom.com/safety. Accessed: 2025-08-26. 3, 4

[5] Vrchat homepage. https://hello.vrchat.com/. Accessed: 2025-08-26. 1, 2, 3, 4, 6

[6] Vrchat safety page. https://help.vrchat.com/hc/en-us/articles/333028197555 39-Safety-Resources-For-Players. Accessed: 2025-08-26. 3, 4

[7] B. S. Abhinaya, A. Sabir, and A. Das. Enabling Developers, Protecting Users: Investigating Harassment and Safety in VR, Mar. 2024. doi: 10.48550/arXiv.2403.05499 1, 2, 8, 9

[8] O. A. Adefope and F. Bayraktar. Social identity framework and gender-based harassment in digital gaming spaces: A scoping review. *Human Behavior and Emerging Technologies*, 2025(1):1811677, 2025. doi: 10.1155/hbe2/1811677 2

[9] F. Argelaguet and C. Andujar. A survey of 3d object selection techniques for virtual environments. *Computers Graphics*, 37(3):121–136, 2013. doi: 10.1016/j.cag.2012.12.003 6, 7

[10] J. N. Bailenson and J. Blascovich. *Encyclopedia of Human-Computer Interaction*, chap. Avatars. Berkshire Publishing Group, 2004. 5

[11] L. Blackwell, N. Ellison, N. Elliott-Deflo, and R. Schwartz. Harassment in social virtual reality: Challenges for platform governance. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), article no. 100, 25 pages, 2019. doi: 10.1145/3359202 1, 2, 4, 5

[12] D. A. Bowman and L. F. Hodges. Formalizing the design, evaluation, and application of interaction techniques for immersive virtual environments. *Journal of Visual Languages & Computing*, 10(1):37–53, 1999. doi: 10.1006/jvlc.1998.0111 1, 3, 4, 6, 7, 8

[13] Q. Chen, J. Cai, and G. Jacucci. "People are Way too Obsessed with Rank": Trust System in Social Virtual Reality. *Computer Supported Cooperative Work (CSCW)*, May 2024. doi: 10.1007/s10606-024-09498-7 1, 4, 6, 9

[14] Q. Chen, S. M. Mousavi, G. Riccardi, and G. Jacucci. Investigating the Use and Perception of Blocking Feature in Social Virtual Reality Spaces: A Study on Discussion Forums. *Proceedings of the ACM on Human-Computer Interaction*, 9(2):1–19, May 2025. doi: 10.1145/3711018 1, 5, 8, 9

[15] Q. Chen, M. M. Spapé, and G. Jacucci. Understanding Phantom Tactile Sensation on Commercially Available Social Virtual Reality Platforms. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–22, Apr. 2024. doi: 10.1145/3637418 1

[16] Q. Chen, Q. Wu, and G. Jacucci. Democratic moderation: Exploring the use and perception of votekicking in social virtual reality. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, article no. 1238, 18 pages. Association for Computing Machinery, New York, NY, USA, 2025. doi: 10.1145/3706598.3713577 1, 7

[17] J. Feine, U. Gnewuch, S. Morana, and A. Maedche. A Taxonomy of Social Cues for Conversational Agents. *International Journal of Human-Computer Studies*, 132:138–161, Dec. 2019. doi: 10.1016/j.ijhcs.2019.07.009 4, 5, 8

[18] C. Fiani, R. Bretin, M. McGill, and M. Khamis. "abracadabra, block the bullies!": Child perceptions of manual, semi-automated and automated interventions against group harassment in social vr. In *Proceedings of the 24th Interaction Design and Children*, IDC '25, 26 pages, p. 314–339. Association for Computing Machinery, New York, NY, USA, 2025. doi: 10.1145/3713043.3728850 2, 5, 8

[19] G. Freeman and D. Acena. Hugging from a distance: Building interpersonal relationships in social virtual reality. In *Proceedings of the 2021 ACM International Conference on Interactive Media Experiences*, IMX '21, 12 pages, p. 84–95. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3452918.3458805 1

[20] G. Freeman and D. Acena. "acting out" queer identity: The embodied visibility in social virtual reality. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2), article no. 263, 32 pages, Nov. 2022. doi: 10.1145/3555153 2

[21] G. Freeman, L. Li, and K. Schulenberg. "i have abused someone who abused me": Understanding people who have experienced both sides of harassment accusations in social vr. *Proc. ACM Hum.-Comput. Interact.*, 9(2), article no. CSCW107, 26 pages, May 2025. doi: 10.1145/3711005 1, 2, 5

[22] G. Freeman and D. Maloney. Body, avatar, and me: The presentation and perception of self in social virtual reality. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW3), article no. 239, 27 pages, Jan. 2021. doi: 10.1145/3432938 1

[23] G. Freeman, D. Maloney, D. Acena, and C. Barwulor. (Re)discovering the Physical Body Online: Strategies and Challenges to Approach Non-Cisgender Identity in Social Virtual Reality. In *CHI Conference on Human Factors in Computing Systems*, pp. 1–15. ACM, New Orleans LA USA, Apr. 2022. doi: 10.1145/3491102.3502082 2

[24] G. Freeman, S. Zamanifard, D. Maloney, and D. Acena. Disturbing the Peace: Experiencing and Mitigating Emerging Harassment in Social Virtual Reality. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–30, Mar. 2022. doi: 10.1145/3512932 1, 2

[25] G. Freeman, S. Zamanifard, D. Maloney, and A. Adkins. My body, my avatar: How people perceive their avatars in social virtual reality. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, 8 pages, p. 1–8. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3334480.3382923 1

[26] R. Jakobson. Essais de linguistique générale. *Les Etudes Philosophiques*, 18(4), 1963. 1, 4, 8

[27] B. Kordyaka, K. Jahn, and B. Niehaves. Towards a unified theory of toxic behavior in video games. *Internet Research*, 30(4):1081–1102, June 2020. doi: 10.1108/INTR-08-2019-0343 2

[28] Z. Liao, H. Zhao, A. Kulkarni, S. Chattrath, and A. X. Zhang. Building proactive and instant-reactive safety designs to address harassment in social virtual reality. *Proc. ACM Hum.-Comput. Interact.*, 9(7), article no. CSCW279, 38 pages, Oct. 2025. doi: 10.1145/3757460 2, 6, 9

[29] Y. Liu, C. K. Yiu, Z. Zhao, W. Park, R. Shi, X. Huang, Y. Zeng, K. Wang, T. H. Wong, S. Jia, J. Zhou, Z. Gao, L. Zhao, K. Yao, J. Li, C. Sha, Y. Gao, G. Zhao, Y. Huang, D. Li, Q. Guo, Y. Li, and X. Yu. Soft, miniaturized, wireless olfactory interface for virtual reality. *Nature Communications*, 14(1):2297, May 2023. doi: 10.1038/s41467-023-37678-4 1

[30] C. MacArthur, E. Kukshinov, D. Harley, T. Pawar, N. Modi, and L. E. Nacke. Experiential disparities in social VR: Uncovering power dynamics and inequality. *Frontiers in Virtual Reality*, 5:1351794, Aug. 2024. doi: 10.3389/frvir.2024.1351794 2

[31] D. Maloney, G. Freeman, and A. Robb. Stay Connected in An Immersive World: Why Teenagers Engage in Social Virtual Reality. In *Interaction Design and Children*, pp. 69–79. ACM, Athens Greece, June 2021. doi: 10.1145/3459990.3460703 9

[32] T. Marwa, O. Salima, and M. Souham. Deep learning for online harassment detection in tweets. In *2018 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, pp. 1–5. IEEE, Tebessa, Oct. 2018. doi: 10.1109/PAIS.2018.8598530 8

[33] J. McVeigh-Schultz, A. Kolesnichenko, and K. Isbister. Shaping pro-social interaction in vr: An emerging design framework. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, 12 pages, p. 1–12. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3290605.3300794 1

[34] J. McVeigh-Schultz, E. Márquez Segura, N. Merrill, and K. Isbister. What's it mean to "be social" in vr? mapping the social vr design ecology. In *Proceedings of the 2018 ACM Conference Companion Publication on Designing Interactive Systems*, DIS '18 Companion, 6 pages, p. 289–294. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3197391.3205451 1

[35] M. Mozafari, R. Farahbakhsh, and N. Crespi. A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media, Oct. 2019. doi: 10.48550/arXiv.1910.12574 8

[36] R. C. Nickerson, U. Varshney, and J. Muntermann. A method for taxonomy

development and its application in information systems. *European Journal of Information Systems*, 22(3):336–359, May 2013. doi: 10.1057/ejis.2012.26 1, 3

[37] J. Outlaw and B. Duckles. Why women don't like social virtual reality : A study of safety, usability and self-expression in social vr. *The Extended Mind*, 2017. 2, 5

[38] D. Picard. De la communication à l'interaction : l'évolution des modèles. *Communication et langages*, 93(1):69–83, 1992. doi: 10.3406/colan.1992.2380 1, 4, 8

[39] E. Reid, R. L. Mandryk, N. A. Beres, M. Klarkowski, and J. Frommel. Feeling Good and In Control: In-game Tools to Support Targets of Toxicity. *Proceedings of the ACM on Human-Computer Interaction*, 6(CHI PLAY):1–27, Oct. 2022. doi: 10.1145/3549498 9

[40] R. S. Herz. Olfactory Virtual Reality: A New Frontier in the Treatment and Prevention of Posttraumatic Stress Disorder. *Brain Sciences*, 11(8):1070, Aug. 2021. doi: 10.3390/brainsci11081070 1

[41] K. Schulenberg, G. Freeman, L. Li, and C. Barwulor. "creepy towards my avatar body, creepy towards my body": How women experience and manage harassment risks in social virtual reality. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2), article no. 236, 29 pages, 2023. doi: 10.1145/3610027 1, 2, 5

[42] K. Schulenberg, G. Freeman, L. Li, and B. P. Knijnenburg. Does who you are or appear to be matter?: Understanding identity-based harassment in social vr through the lens of (mis)perceived identity revelation. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW2), article no. 384, 40 pages, Nov. 2024. doi: 10.1145/3686923 2

[43] K. Schulenberg, L. Li, C. Lancaster, D. Zytko, and G. Freeman. "We Don't Want a Bird Cage, We Want Guardrails": Understanding & Designing for Preventing Interpersonal Harm in Social VR through the Lens of Consent. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–30, Sept. 2023. doi: 10.1145/3610172 1, 2, 4, 9

[44] W. Y. Tang, F. Reer, and T. Quandt. Investigating sexual harassment in online video games: How personality and context factors are related to toxic sexual behaviors against fellow players. *Aggressive Behavior*, 46(1):127–135, Jan. 2020. doi: 10.1002/ab.21873 2

[45] J. Tschanter, C. Merz, C. Wienrich, and M. E. Latoschik. Towards understanding harassment in social virtual reality: A study design on the impact of avatar self-similarity. In *2025 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 172–177, 2025. doi: 10.1109/VRW66409.2025.00043 2, 5

[46] W.-J. Tseng, E. Bonnail, M. McGill, M. Khamis, E. Lecolinet, S. Huron, and J. Gugenheimer. The dark side of perceptual manipulations in virtual reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, article no. 612, 15 pages. Association for Computing Machinery, New York, NY, USA, 2022. doi: 10.1145/3491102.3517728 5

[47] N. Vishwamitra, H. Hu, F. Luo, and L. Cheng. Towards Understanding and Detecting Cyberbullying in Real-world Images. In *Proceedings 2021 Network and Distributed System Security Symposium*. Internet Society, Virtual, 2021. doi: 10.14722/ndss.2021.24260 8

[48] E. Volchan, G. G. Souza, C. M. Franklin, C. E. Norte, V. Rocha-Rego, J. M. Oliveira, I. A. David, M. V. Mendlowicz, E. S. F. Coutinho, A. Fiszman, W. Berger, C. Marques-Portella, and I. Figueira. Is there tonic immobility in humans? Biological evidence from victims of traumatic stress. *Biological Psychology*, 88(1):13–19, Sept. 2011. doi: 10.1016/j.biopsycho.2011.06.002 9

[49] M. Wang, W.-T. Shu, Y.-J. Li, and W. Li. Can I Get There? Negotiated User-to-User Teleportations in Social VR. *IEEE Transactions on Visualization and Computer Graphics*, 31(5):2320–2330, May 2025. doi: 10.1109/TVCG.2025.3549572 2

[50] N. Wang, J. Zhou, J. Li, B. Han, F. Li, and S. Chen. HardenVR: Harassment Detection in Social Virtual Reality. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 94–104. IEEE, Orlando, FL, USA, Mar. 2024. doi: 10.1109/VR58804.2024.00033 8

[51] M. Weerasinghe, S. Macdonald, C. Fiani, J. O'Hagan, M. Chollet, M. McGill, and M. Khamis. Beyond mute and block: Adoption and effectiveness of safety tools in social vr, from ubiquitous harassment to social sculpting. *IEEE Transactions on Visualization and Computer Graphics*, 31(5):3275–3284, 2025. doi: 10.1109/TVCG.2025.3549860 1, 2, 3, 4, 5, 8

[52] X. Wei, X. Jin, G. Lin Kan, Y. Yan, and M. Fan. Systematic Literature Review of Using Virtual Reality as a Social Platform in HCI Community. *Proceedings of the ACM on Human-Computer Interaction*, 9(2):1–36, May 2025. doi: 10.1145/3711078 1

[53] Q. Zheng, S. Xu, L. Wang, Y. Tang, R. C. Salvi, G. Freeman, and Y. Huang. Understanding safety risks and safety design in social vr environments. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW1), article no. 154, 37 pages, 2023. doi: 10.1145/3579630 1, 2, 3, 4, 8, 9

[54] D. Zytko and J. Chan. The Dating Metaverse: Why We Need to Design for Consent in Social VR. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2489–2498, May 2023. doi: 10.1109/TVCG.2023.3247065 2