IN[The Hate Booth]: a Gamified Installation to Counteract Hate Speech

Susana Costa[1](✉), Mirian Tavares[1], José Bidarra[2], and Bruno Mendes da Silva[1]

[1] Research Centre for Arts and Communication, Algarve University, Faro, Portugal
{srsilva,mtavares,bsilva}@ualg.pt
[2] Research Centre for Arts and Communication, Aberta University, Edmonton, Canada
jose.bidarra@ualb.pt

Abstract. Playing an online game, interacting on a social network or in a digital gaming community is part of the daily lives of most children and youth, with effects on the development of the personality, influence on the behavior and on the ability to manage conflicts.

Studies and reports have been analyzing the interactions of online players, on gaming platforms and communities, as consumers and content producers, with the aim of understanding and finding effective ways to prevent hate speech from proliferating in these digital environments.

In this article, we present a gamified installation, combining narrative and participatory approaches, as a response to the proliferation of online hate speech. The game-installation [*IN]The Hate Booth* consists of a light booth, where the interactor can find an interactive fiction game inspired by the videogame universe. This game will be the basis of a pedagogical itinerary, aiming to reflect on experiences with online hate speech and its effects inside and outside the virtual world [1, 2].

The initiative seeks to contribute to achieving and developing the sixteenth goal of the United Nations 2030 Agenda for Sustainable Development, Peace, Justice and Strong Institutions: the construction of peaceful and just societies, and effective, accountable and inclusive democracies at all levels.

Keywords: Online Hate Speech · Counternarratives · Gamification

1 Introduction

The Internet, and mostly the social networks, transformed the way individuals communicate. From an era in which ideas were transmitted in a slower and limited way, we evolved to the era of superinformation, based on countless diffusion channels to interact, share and express ideas and information, instantly and massively.

The media have been the stage for strong discussions that, many times, result in the use of offensive and discriminatory language [3, 4]. The use and dissemination of hate speech permeates online platforms. As a result, hate speech has been recognized as a

A. L. Brooks (Ed.): ArtsIT 2022, LNICST 479, pp. 161–173, 2023.
https://doi.org/10.1007/978-3-031-28993-4_12

serious problem, giving rise to several international initiatives as countermeasures to the problem. At the end of the last decade, the academic interest in hate speech had a significant increase, reflected, for example, in the volume of production indexed in Web of Science (WoS), which increased from 42 to 162 between 2013 and 2018 [2–4].

Researchers have been describing online hate speech as a set of behaviors categorized as toxic in relation to constantly renegotiating and evolving social norms [2]. Sellars [8] surveyed a set of common traits that help to identify hate speech: the fact of addressing a group or an individual, as a member of a group; the presence of content that expresses hate and may cause harm; the intention to harm; the public nature of the discourse and, finally, the existence of a context that makes the violent response possible. From these common traits, different types of hate speech can be identified, motivated by gender, sexual identity, nationality, historical events, or religious beliefs [2, 3, 9–11].

Citron and Norton [13] define and analyze four forms of response to online hate messages: (2) inaction (3) exclusion/suspension of messages and users (4) education (5) counter-narratives. The first and second types of reactions have been the most used by large technology corporations. According to the authors, the silence in response to digital hate also carries significant expressive costs, so it can contribute to the legitimacy of a type of hate speech. In turn, blocking the users of hate messages may work as a short-term solution, on an individual scale, but it continues to be detrimental to the community, especially from the point of view of freedom of expression.

Available data related to these experiences is often difficult to assess as it is in private databases. One of the most recent datasets, a report focusing on the analysis of cyber hate experiences of children aged 11–17 years in 10 European countries concluded that exposure to hate speech increases with age, a trend likely correlated with greater overall engagement with the virtual world [2, 14, 16].

Hateful messages need to be limited, because they can violate the dignity of the others; however, doing so often creates a conflict between fundamental rights: on the one hand, freedom of expression and, on the other, the right to equality, inclusion, and protection. The complex balance between these rights has been the subject of analysis when addressing the problem of hate speech on the Internet.

2 Previous Research on Hate Speech in Video Games and Game Platforms

Online games provide players with the emotion of competition [5] and the social interactions that are provided can be positive or negative. Many gamers believe online games are a positive experience and a unique and privileged way to keep in touch with friends or create new networks of contacts. These positive aspects play an important role in the community in the digital and physical lives of the players [19].

The negative aspects, many times provided by the competition and experience of the game, can lead to verbal expression of profanity and obscenity, tolerated as a common reaction in moments of anger and frustration [13, 20, 21].

During the games, the interaction in chats is common, and the dialogues diverge between praise and negative or ironic comments about the performance of the game, personal insults based on sexual orientation or ethnicity, situations of harassment and attacks on minorities [22]. The hate speech in digital games is frequently the result of these interactive dynamics between gamers, in unmoderated activities, such as the formation of teams, the discussion of strategies in chats or livestreaming on game platforms and communities, which are commonplace to spread this type of toxic behavior [23]. Resorting to censorship as a response to these manifestations of hate can, at times, oppose freedom of expression, the pillar and conquest of democratic societies, the foundation of self-realization, autonomy, democracy and truth [24].

Social networks have been monitoring hate speech by identifying and reporting content to platform creators and administrators. When we analyze this type of phenomenon in the field of videogames, we deal with different characteristics, such as the access and the less public nature of these spaces, when compared to *Twitter*, *Tiktok*, *Youtube*, or *Facebook*.

In 2017, when more than 500 white supremacists marched in Charlottesville, in the United States of America, they showed the world its organizational capacity (programming, logistics and transportation), all set up in *Discord* chat rooms, a communication tool initially created for the gaming community, which allows the creation of chat rooms and groups to unite gamers. Following the incidents in Charlottesville, where there were fatalities, this platform imposed restrictive measures on hate speech, by banning several users who express sympathy or connection to neo-Nazi or white supremacist ideologies and by prohibiting, through censorship, messages of harassment or threats.

In 2019, following the attack on the Christchurch Mosque, the author, Brenton Tarrant, stated that *Facebook*, *Reddit* and *Youtube* had been a significant source of information and inspiration [25]. According to Lamphere-Englund and Bunmathong [26], prominent social networks have intensified the mediation of extremist content, pushing this type of expression to new forums, namely games and gaming sites. The authors illustrate their conclusions with several examples that attest to the links between gaming and extremism.

Steam, the community of gamers and the store, refused to block games or content in defense of the right to freedom of expression, reaffirming itself as a market for games closed to cultural disputes. *Twitch* and *YouTube* are other platforms that allow to attend live broadcasts of everything, including games.

European legislation, as well as some national laws, have taken important steps in combating hate speech online, for example through the establishment of Codes of Conduct, the updating of legislation on this matter, or criminal sanctions. On the other hand, there have been some initiatives by the most prominent technological corporations in response to this problem, through user policies that converge on blocking individuals, assuming a commitment to act quickly in case of complaints about this type of abuse: "Despite initial resistance, and following public pressure, some of the companies owning these spaces have become more responsive towards tackling the problem of hate speech online, although they have not (yet) been fully incorporated into global debates" [21].

Previous studies point to game design as one of the primary influencers of user behavior, having identified a correlation between a more competitive type of game and the proliferation of hate speech [2, 4–6, 17, 28, 29]. In turn, dissociative anonymity and imagination, invisibility, asynchrony, minimization of state and authority, individual differences and predisposition may trigger toxic disinhibition, as defined by Suler [5], also contributing to the proliferation of the phenomenon. Suler proposes that the effect of anonymity on the Internet leads individuals to a feeling of freedom, leading to actions different from those they would have if they were face to face with the other. This phenomenon favors the proliferation of trolls and the use of hate speech, characterised by the demonstration of power or expression of frustration in the face of defeat. This behaviour is sometimes detrimental to the physical and psychological well-being as well as the self-esteem of aggressors and victims [2, 19, 28, 29].

Analyzing games and platforms user policies, it is possible to recognize that most of them has a defined policy in relation to hate speech [19, 22, 27] in which the terms of use establish silencing sanctions. The ways of identifying this type of discourse also follow similar protocols in games and platforms: use of Artificial Intelligence (AI), tuned through "machine learning" or "human-in-the-loop" processes (AI systems trained and fine-tuned by humans to accurately recognize online hate speech), associated with teams of varying sizes that work systematically to detect these phenomena. The game *Roblox*, for example, claimed to have a team of more than 3000 individuals for this purpose, however during the last coronavirus pandemic there were several reports of attacks on players, under the age of 16, with African-American avatars inside the game [12, 25, 32, 33].

Attention given to online hate speech, particularly in videogames and gaming communities, is a reaction to the dissemination of this type of speech and the need to guarantee a safe environment in digital spaces. The approaches to the problem are addressed mainly by two types of strategies for containment and prevention: (1) automatic detection and classification, based on AI tools and computer science methodologies; (2) the construction and dissemination of counter-narratives.

Within the game's communication channels, some toxic communication is fueled by reactions to in-game events, player performance, and other competitive features. The widespread use of insults in video games is fueled by the same anonymity that covers abuse in cyberspace, and game creators are not fully able to predict the consequences that each game may have. On the other hand, in addition to this spontaneous discourse, the organized use of platforms as methodologically planned training and radicalization rooms is also manifest.

2.1 Don't Feed the Troll

"Don't feed or troll" is a popular axiom among Internet users, emphasizing that indifference is the best, and perhaps the only, way to respond to trolls.

In the last decade, trolls and bots have entered the world vocabulary to unravel phenomena that have been constituted, above all, as a challenge for the users and managers of social platforms, because they have implications in the management and governance of online communities. Trolling is regularly prevalent on all social networks and is related to the proliferation of hate speech. Aided by the easy ways to creating anonymous or

false online profiles and by the atomized nature of interaction on the Internet, it can represent a threat, when it systematically reaches minority groups [31, 34].

Previous studies confirm that both trolls and bots exert significant influence in driving digital toxicity. Bots can spread hate online, building groups and supporting opinion leaders who promote a certain segregation. Stella et al. [35], Uyheng et al. [22] and Robles et al. [36] correlates the proliferation of negative and inflammatory narratives, spread by trolls and bots, in political or social events with a polarization of opinions, intolerance and the increase in hate speech online.

The performances of trolls demand an understanding of two idiosyncratic elements of a community. Cruz et al. [34] showed that transgression is just one element of a constellation of complementary practices – learning, assimilation and transgression – that constitute trolling and that require a degree of knowledge and commitment to a given community.

Due to the increase in the incidence of trolling in online gaming environments, several studies have emerged to analyze this trend. Thacker Griffiths [37] examined the frequency, motivation, and effects of trolling in video games. The results showed that the trolls tended to keep gaming sessions longer, and that the types of trolling mostly included sexism and racism. The main reasons given for this practice include fun, teasing and vindictiveness. On the other hand, the study shows that being identified as a troll is positively associated with an increase in self-esteem, while being targeted by trolling is perceived as rather negative, linked to a decrease in self-esteem. In turn, Cook et al. [38] confirm that trolls do not have a uniform behavior and motivation, distinguishing three categories of action (1) attack, (2) seeking sensations and (3) seeking interactions. The first seek to inflict some kind of pain on their victims, the second seek new emotions, and the third seek attention through new interactions. This analysis allows us to conclude that trolling is characterized by its instrumentality, and not by an arbitrary nature.

Finally, Cook et al. [38] highlight the importance of a sense of community, in which the trolls are aware of one another and, at times, undertake group movements. These events even though considered normal and unavoidable, are seen as deviant and constitute negative experiences in the domain of cyberbullying and hate.

3 The Methodology: Counter-Narratives to Hate Speech

One of the responses indicated by Citron and Norton [13] to online hate speech is the use of counter-narratives to copy with the phenomena. A counter-narrative is a message that offers a positive alternative to extremist, racist, xenophobic, or any other propaganda that affects individual freedom. It is a way to deconstruct or delegitimize a certain type of discourse that affects the dignity of the other.

Tuck and Silverman [39] suggest that to create effective counter-narratives, it is necessary to consider factors such as age and language, offering content capable of generating thoughts, feelings, memories, and reflections. The authors argue that the creation of counter-narrative content can be a slow process, which also requires the expansion, redirection, and recreation of existing content.

To date, a limited number of studies have been carried out on counter-narratives as a response to the massive growth of online hate speech. According to Chung et al. [40], studies have focused on identifying successful counter-narratives [23, 27, 41] evaluating their effectiveness [18, 32, 42, 43] and the characteristics of counter-narratives [44]. In particular, analyzing *Twitter* conversations, Wright et al. [41] show that some discussions between strangers can induce favorable changes in speech and attitudes.

In their studies around the use of counter-narratives to combat online hate speech, Benesch et al. [41] distinguish eight groups of reaction: (1) presenting facts against hate speech; (2) presenting contradictions in hate speech; (3) warning of the offline or online consequences of hate speech; (4) manifesting affiliation with a particular characteristic of the speaker, seeking empathy and deterrence; (5) denunciating hateful speech, through the mechanisms of digital platforms; (6) responding with humor; (7) responding with a positive tone and (8) showing hostility.

In the project *IN[The Hate Booth]* we explore a mixed methodology, based on the creation of counter-narratives, capable of incorporating partnerships, content creation and developing a set of strategies based on gamification, to achieve a measurable and replicable impact.

The concept of culture of convergence [46], which defines the technological, economic, cultural, and social transformations perceived in the contemporary scenario of the media, is the starting point for a participatory approach, considering it the most effective in the development of capabilities and tools for change.

The narrative approach, namely fiction-based research, used to create the counter-narratives, can cause changes in the way individuals relate to themselves and others [47], since the research developed is more truthful, meaningful, useful, accessible, and human.

4 The Game-Installation *IN[The Hate Booth]*

Based on the proposals of Tuck and Silverman [39] and Citron and Norton [13], we combined two axes to generate the installation *IN[The Hate Booth]*: the development of a counter-narrative, and a pedagogical itinerary with an educational purpose.

This game-installation reflects cyberspace as fertile ground for the toxic disinhibition of the hatred discourse. Trolls and bots escape between the watched but unregulated space of the Internet. It is an engaged proposal, based on interactive fiction, developed as a digital installation and as a game.

Currently, young people are increasingly becoming consumers and producers of media, and there is an urgent need to provide them with knowledge and skills that enable to have a more informed level of consumption and media knowledge. The game-installation *IN[The Hate Booth]* seeks to engage the audience as a starting point to address the problem of hate speech and its psychological, social and political consequences.

4.1 The Concept

This Digital Art project is a counter-narrative against online hate speech. It is composed of two dimensions, one physical, a light cabin; and a virtual one, an interactive fiction game that invites the interactor to a path of discovery.

The artifact aims to achieve four effects: a) contribute to the immersion of the interactor - obfuscating what is around while keeping the focus on the game; b) cause obfuscation and discomfort to the interactor, evidencing the effects of hate speech; c) refer, metaphorically, to the sense of stage and role, remembering that everyone has rights and responsibilities in the digital world, as in the physical world, being able to assume their own *persona* in the containment of this phenomenon and, finally, d) it represents virtual communication, inherent to online hate speech, linked to the process of toxic disinhibition.

The cabin is isolated by three panels, inviting the interactor to an immersive experience. Inside the cabin, the interactor finds a screen with the announcement that the feed of a web page will be discontinued, and the only access will correspond to the post-mortem, guaranteeing the documentation of the incident that led to the decision and thanking the regular fans and visitors. The cancellation leads to questions voiced in the comments box: What happened to the authors? Why did they decide to stop writing?

By choosing to find out what happened, it will be possible to find a series of archive messages and, reading and decoding them, the interactor faces the intensification of various types of hate messages among the followers (Fig. 1).

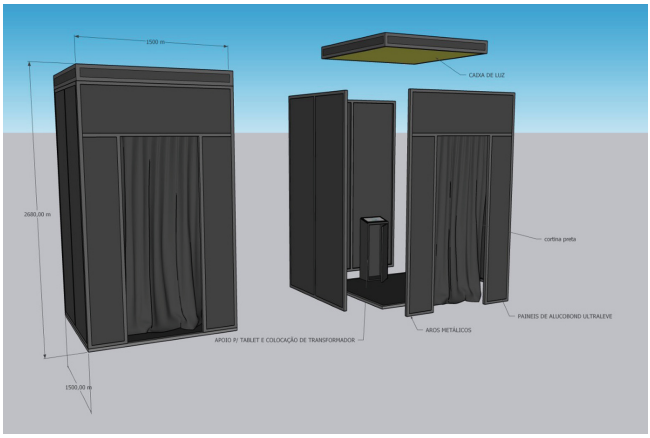

Fig. 1. *IN[The Hate Booth]:* Image of the installation prototype

For the logo of the installation-game, we used the square brackets underlining the experience of the booth: an immersive voyage, isolated from the outside. Inside the installation, a magic circle is created where special rules prevail [45] (Fig. 2 and 3).

The pedagogical itinerary provides the orientation of a training path in which students become co-authors of the interactive installation. In addition to the individual experimentation of the artifact, a collective work will be proposed to produce new clues and new pages that can continue the game, with the purpose of involving students in participation, intervention and an active awakening to the problem. In addition, the installation offers the interactors the possibility to comment on the posts, also involving them in the discussion.

Fig. 2. *IN[The Hate Booth]:* Image of the installation prototype

Fig. 3. *IN[The Hate Booth]:* Image captured in the game environment

4.2 The Technical Features

The script was developed in the application *Celtx* and in *Twinery* to allow the visualization of the possibilities of paths and choices. A *WordPress* domain was created allowing the access of the game with the different pages. To advance the ten levels the interactors must read the posts and the comments to identify a main theme, a keyword, which allows passing to the next post. When the interactor can't find the keyword, a clue is given by underlining in yellow an important passage of the text.

The physical dimension of the installation consists of a black cabin, measuring 1.50 × 1.50 × 2.68 m, divided into 3 panels and 1 roof. Outdoor area in self-adhesive vinyl; closed with a curtain. Lighting will be provided by RGB wallwhashers with a transformer and motion sensor for color change.

The cabin has the following features

- Base in metallic structure coated with black laminate with leveling paters;
- Metal structure coated on the outside with very light aluminum composite;
- Ceiling with opaline acrylic light box with RGB LED lighting;
- The door with a black flannel curtain, fireproof, easy to assemble with ties (allows to enter and prevent the light from going out);

– The lighting will have a timer/sequencer that will be programmed to
– The light color changes every 30 s;
– The lighting will take into account the presence of the interactor, avoiding energy waste when the cabin is empty.

4.3 The Functional Features

The interactor enters the cabin and is faced with an illuminated and closed environment. On the screen is the post-mortem of a video game website. In this post-mortem the game mechanism is also presented. One of the users, in the comments, writes the following:

> "MetaHacker
>
> Guys, I already found out what happened! Type "Welcome" into the search bar and find out what made them close the page…"
>
> And on the second page, you can read:
>
> "Neon
>
> What will be the keyword to get to the next post? Tip: What is the main theme of the blog?"

The interactor understands that based on the theme of a given entry, he can advance, through the search bar, on the site that has been discontinued. So, he can discover and experience the escalation of hate, in the comments, that led to its end. When the interactor cannot find out in three tries what the keyword is, the game gives a clue by putting a part if the text that leads to that keyword in yellow.

In addition to the individual enjoyment of the game inside the booth, a collective work will be proposed, based on a pedagogical itinerary that provides the guidance of a training path in which students become co-authors of the interactive installation. In addition to the individual experimentation of the artifact, a collective work will be proposed to produce new clues and new pages that can continue the game, with the purpose of involving students in participation, intervention and an active awakening to the problem. The installation also offers the interactors the possibility to comment on the posts, involving them in the discussion.

From the experimentation of the installation-game questionnaires will be applied by inquiry and a qualitative data collection will be carried out through focus groups, with young people and adolescents aged between 10 and 14 years to assess the impact of the artifact.

The age group cut was based on previous studies [19], which point to the need to develop media literacy skills at an early age, when children begin to use game platforms with less parental supervision. On the other hand, the development of the game-installation was based on a target audience between this age group.

5 Conclusion

Online hate speech has been taking up more and more space on the Internet, particularly in gaming environments and associated communities. If, on the one hand, the individual

right to freedom of expression is inalienable and indisputable, it is no less important to underline that the exercise of this right implies responsibility and respect for the Other and for difference.

Misogyny, racism, anti-Semitism, homophobia, xenophobia and other forms of alterophobia have various mechanisms for producing victims and causing harm. The concern of democratic governments is precisely the solution to this problem, without harming the values of freedom of expression, seeking a sensible balance between freedom and equality.

The study of the state of the art in this field shows that the path of silencing, application of sanctions and criminalization as a response to hate speech, strategies used by the main social networks such as *Youtube*, *Facebook* or *Discord* have not shown effectiveness in containing the phenomenon, often posing problems in terms of freedom of expression.

The objective of this article was to analyze the problem and survey the solutions that have been found by previous studies and initiatives. Gamification, counter-narratives, and art are the axes we propose in our approach to hate speech, focusing on game culture and social spheres as engines for the promotion of democratic values, critical thinking, and digital citizenship. Among the UN Sustainable Development Goals we find some of the key ideas and concepts basis of the project *IN[The The Booth]*, namely the education for sustainable development and sustainable lifestyles, through the education for human rights, gender equality, and promotion of a culture of peace and non-violence, grounded on global citizenship, and cultural diversity.

Through protocols established with basic schools, the game installation and the itinerary will be experienced by young people between 10 and 14 years old. The experimentation will be accompanied by a workshop and focus group to collect qualitative data about the experience and its contribution to the understanding the phenomenon of hate speech and its input to develop tools and strategies to copy with the problem.

Acknowledgments. This publication is financed by national funds through the project "UIDP/04019/2020 CIAC" of the Foundation for Science and Technology, I.P.

This work is financed by national funds through the FCT – Foundation for Science and Technology, I.P., in the scope of the PhD project integrated in the CIAC: Research Centre for Arts and Communication. Algarve University and Aberta University, UI/BD/150850/2021.

References

1. Costa, S., da Silva, B.M., Tavares, M.: Tackling online hate speech? In: Wölfel, M., Bernhardt, J., Thiel, S. (eds.) ArtsIT, Interactivity and Game Creation (ArtsIT 2021). LNICST, vol. 422, pp. 79–93. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-95531-1_6
2. Costa, S., Tavares, M., Bidarra, J., Mendes da Silva, B.: The *Enredo* game-installation: a proposal to counter hate speech online. In: Martins, N., Brandão, D. (eds.) Advances in Design and Digital Communication III (DIGICOM 2022). Springer Series in Design and Innovation, vol. 27, pp. 307–320. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-20364-0_27
3. Paz, M.A., Montero-Díaz, J., Moreno-Delgado, A.: Hate speech: a systematized review. Sage Open **10**(4), 2158244020973022 (2020). https://doi.org/10.1177/2158244020973022

4. Tontodimamma, A., Nissi, E., Sarra, A., Fontanella, L.: Thirty years of research into hate speech: topics of interest and their evolution. Scientometrics **126**(1), 157–179 (2020). https://doi.org/10.1007/s11192-020-03737-6
5. Kwak, H., Blackburn, J.: Linguistic analysis of toxic behavior in an online video game. In: Aiello, L.M., McFarland, D. (eds.) SocInfo 2014. LNCS, vol. 8852, pp. 209–217. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-15168-7_26
6. Suler, J.: The online disinhibition effect. Cyberpsychology Behav. Impact Internet Multimedia Virtual Reality Behav. Soc. **7**(3), 321–326 (2004). https://doi.org/10.1089/1094931041291295
7. Eur-lex page: Framework decision on combating certain forms and expressions of racism and xenophobia by means of criminal law. https://bit.ly/2WdtFwx. Accessed 18 June 2022
8. Sellars, A.: Defining Hate Speech: Berkman Klein Center Research Publication No. 2016-20, pp. 16–48, Boston University School of Law (2016). https://doi.org/10.2139/ssrn.2882244
9. Blaya, C.: Cyberhate: a review and content analysis of intervention strategies. Aggress. Violent. Behav. **45**, 163–172 (2019). https://doi.org/10.1016/j.avb.2018.05.006
10. Agustina, J.R., Montiel, I., Gámez-Guadix, M.: Cibercriminología y Victimización Online. Editorial Sintesis, Madrid (2020)
11. Deslauriers, P., St-Martin, L., Bonenfant, M.: Assessing toxic behaviour in dead by daylight: perceptions and factors of toxicity according to the game's official subreddit contributors. Game Stud. **20**(4) (2020). https://bit.ly/3hZ37a
12. Hate Base: https://hatebase.org/news/2019/09/27/this-year-companies-will-spend-124b. Accessed 18 June 2022
13. Citron, D.K., Norton, H.: Intermediaries and hate speech: fostering digital citizenship for our information age. BUL Rev. **91**, 1435 (2011). https://ssrn.com/abstract=1764004
14. Council of Europe: Developing media literacy and critical thinking through education and training - council conclusions. https://bit.ly/30zFMqk. Accessed 31 Jan 2022
15. Directorate-General for Justice and Consumers (Ed.): 5th evaluation of the code of conduct (2021). https://bit.ly/3hV0ExX
16. Machackova, H., Blaya, C., Bedrosova, M., Smahel, D., Staksrud, E.: Children's experiences with cyberhate. EU Kids Online (2020). https://doi.org/10.21953/lse.zenkg9xw6pua
17. Sachs, J., Schmidt-Traub, G., Kroll, C., Lafortune, G., Fuller, G., Woelm, F.: The Sustainable Development Goals and COVID-19: Sustainable Development Report 2020. Cambridge University Press (2020). https://bit.ly/3rrQQQH
18. Wachs, S., Wettstein, A., Bilz, L., Gámez-Guadix, M.: Adolescents' motivations to perpetrate hate speech and links with social norms. Comunicar **71**, 9–20 (2022). https://doi.org/10.3916/C71-2022-01
19. Play Your Role: Research Report (2020). https://www.playyourrole.eu/wp-content/uploads/2020/07/PYR-research-report.pdf
20. Breuer, J.: Hate Speech in Online Games. In: Kaspar, K., Gräßer, L. (eds.) Online Hate Speech, pp. 107–112. Kopaed, Düsseldorf (2017)
21. Gagliardone, I., Gal, D., Alves, T., Martinez, G.: Countering Online Hate Speech. United Nations Educational, Scientific and Cultural Organization, Paris (2015)
22. Uyheng, J., Carley, K.M.: Characterizing network dynamics of online hate communities around the COVID-19 pandemic. Appl. Netw. Sci. **6**(1), 1–21 (2021). https://doi.org/10.1007/s41109-021-00362-x
23. Matamoros-Fernández, A.: Platformed racism: the mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. Inf. Commun. Soc. **20**(6), 930–946 (2017). https://doi.org/10.1080/1369118X.2017.1293130
24. Greenawalt, K.: Rationales for freedom of speech. In: Moore, A.D. (ed.) Information Ethics: Privacy, Property, and Power, pp. 278–296. Washington University Press, Washington (2005)

25. Wired: https://www.wired.com/story/roblox-online-games-irl-fascism-roman-empire/. Accessed 23 June 2022
26. Lamphere-Englund, G., Bunmathong, L.: State of Play on Gaming Extremism - An Annotated Bibliography. The Extremism and Gaming Research Network (2021)
27. Hokka, J.: PewDiePie, Racism and Youtube's neoliberalist interpretation of freedom of speech. Convergence **27**(1), 142–160 (2021). https://doi.org/10.1177/1354856520938602
28. Soral, W., Bilewicz, M., Winiewski, M.: Exposure to hate speech increases prejudice through desensitization. Aggress. Behav. **44**(2), 136–146 (2018). https://doi.org/10.1002/ab.21737
29. Harriman, N., Shortland, N., Su, M., Cote, T., Testa, M.A., Savoia, E.: Youth exposure to hate in the online space: an exploratory analysis. Int. J. Environ. Res. Public Health **17**, 8531 (2020). https://doi.org/10.3390/ijerph17228531
30. Arbeau, K., Thorpe, C., Stinson, M., Budlong, B., Wolff, J.: The meaning of the experience of being an online video game player. Comput. Hum. Behav. Rep. **2**, 100013 (2020). https://doi.org/10.1016/j.chbr.2020.100013
31. Brockmyer, J.: Media violence, desensitization, and psychological engagement. In: The Oxford Handbook of Media Psychology, vol. 24(1), pp. 212–222 (2013).https://doi.org/10.1016/j.chc.2014.08.001
32. Funk, J.B., Baldacci, H.B., Pasold, T., Baumgardner, J.: Violence exposure in real-life, video games, television, movies, and the internet: is there desensitization? J. Adolesc. **27**, 23–39 (2004). https://doi.org/10.1016/j.adolescence.2003.10.005
33. The New York Times: //www.nytimes.com/2021/09/08/parenting/online-hate-groups-kids.html. Accessed 18 June 2022
34. Cruz, A., Seo, Y., Rex, M.: Trolling in online communities: a practice-based theoretical perspective. Inf. Soc. **34**(1), 15–26 (2018). https://doi.org/10.1080/01972243.2017.1391909
35. Stella, M., Ferrara, E., De-Domenico, M.: Bots increase exposure to negative and inflammatory content in online social systems. In: Kleinberg, J. (ed.) Proceedings of the National Academy of Sciences, vol. 115, pp. 12435–12440 (2018). https://doi.org/10.1073/pnas.1803470115
36. Robles, J., Guevara, J., Casas-Mas, B., Gómez, D.: When negativity is the fuel. Bots and political polarization in the COVID-19 debate. Comunicar **71**, 63–75 (2022). https://doi.org/10.3916/C71-2022-05
37. Thacker, S., Griffiths, M.D.: An exploratory study of trolling in online video gaming. Int. J. Cyber Behav. Psychol. Learn. **2**(4), 17–33 (2012). https://doi.org/10.4018/ijcbpl.2012100102
38. Cook, C., Schaafsma, J., Antheunis, M.: Under the bridge: an in-depth examination of online trolling in the gaming context. New Media Soc. **20**(9), 3323–3340 (2018). https://doi.org/10.1177/1461444817748578
39. Tuck, H., Silverman, T.: The Counter-Narrative Handbook. Institute for Strategic Dialogue (2016)
40. Chung, Y., Tekiroglu, S., Guerini, M.: CONAN--COunter NArratives through Nichesourcing: a multilingual dataset of responses to fight online hate speech. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2819–2829 (2019). https://aclanthology.org/P19-1271/
41. Wright, L., Ruths, D., Dillon, S.H., Benesch, S.: Vectors for counterspeech on Twitter. In: Proceedings of the First Workshop on Abusive Language Online, pp. 57–62 (2017). https://aclanthology.org/W17-3009.pdf
42. Schieb, C., Preuss, M.: Governing hate speech by means of counterspeech on Facebook. In: 66th ICA Annual Conference, pp. 1–23 (2016). https://www.researchgate.net/publication/303497937_Governing_hate_speech_by_means_of_counterspeech_on_Facebook
43. Ernst, J., et al.: Hate beneath the counter speech? A qualitative content analysis of user comments on Youtube related to counter speech videos. J. Deradicalization (10), 1–49 (2017). https://journals.sfu.ca/jd/index.php/jd/article/view/91/80

44. Mathew, B., et al.: Thou shalt not hate: countering online hate speech. Web Soc. Media **13**(01), 369–380 (2018). https://ojs.aaai.org/index.php/ICWSM/article/view/3237
45. Huizinga, J.: Homo Ludens: um estudo sobre o elemento lúdico na cultura. Edições 70, Lisboa (2003)
46. Jenkins, H.: Cultura da Convergência. Aleph, São Paulo (2003)
47. Gubrium, A., Harper, K.: Participatory Visual and Digital Methods. Le Coast Press, Londres (2013)
48. Clandinin, D.J., Rosiek, J.: Mapping a landscape of narrative inquiry: borderland spaces and tensions. In: Clandinin, D.J. (ed.) Handbook of Narrative Inquiry: Mapping a Methodology, pp. 35–75. Sage Publishing (2007). https://bit.ly/3sjNXSP