Take this.

PERCEPTIONS OF HARM IN ONLINE GAMING

Insights from players and industry professionals

Rachel Kowert, PhD
Elizabeth D. Kilmer, PhD

AUGUST 2024

01

Executive Summary

- It is unclear whether the high prevalence of harmful behaviors in gaming spaces may be, at least partially, due to a lack of consensus as to what is, and what is not, considered harmful within these communities.
- To assess whether or not player and industry groups are aligned in their perceptions of harm in gaming spaces, an online survey was conducted and administered to player communities and game industry professionals who specialize in player safety.
- Differences in perceived prevalence and severity of harms in gaming spaces were found between players and industry professionals.
- Results suggest that player and professional groups are generally aligned in their perception of the severity of in-game harms; however, industry professionals report online harms in gaming spaces as more prevalent.
- Differences in perception of prevalence may be indicative of more consistent exposure to these harms as professionals in the industry or indicate that industry groups have been effective at shielding players from a significant number of harms across categories.
- The greatest disparities in prevalence are found for the most severe social offenses, including CSAM, doxxing, hate speech, incitement of violence and swatting, suggesting trust and safety professionals are effectively shielding players from a significant number of these offenses.

02

Introduction

The prevalence and harmful nature of toxicity has been well documented within gaming and game-adjacent spaces. A majority of young people and adult players report experiencing harassment in online games (ADL, 2024). As many as eight out of ten players report being a victim of some form of toxic behavior in games, with nine out of ten of these players reporting that these behaviors have some negative impact on their mental health (Kowert, Kilmer, & Newhouse, 2024). Perhaps more troubling, is the fact that the overall percentage of players who report witnessing or experiencing toxic behavior is on the rise (Unity, 2023).

There are many reasons as to why the prevalence of harmful and so-called "toxic" behavior is so prevalent in gaming spaces. People feel emboldened to be hurtful under the guise of anonymity and invisibility (Suler, 2004). There is also a cultural element to consider, with these kinds of behaviors becoming culturally "normalized" in gaming spaces specifically (Beres et al, 2021; Kowert, Kilmer, Newhouse, in press). The growing prevalence of these behaviors in gaming spaces is also thought to be due to an ineffectiveness in moderation at scale and lack of prioritization for player safety.

Generally speaking, moderation at scale is a challenge for any social platform. Games are no exception. The sheer volume of data, delay in moderation actions, impact of disruptive players on human moderators, and ambiguity about applying the "right" action, are all continued challenges within the space (Kocielnik et al., 2024) that make moderation at scale inefficient at best and ineffective at worst. One particular challenge in this space is the potential misalignment across player and professional communities about what behaviors should be actioned at all. What is considered "harmful" can vary from person to person and platform to platform. This relative subjectivity in what is considered toxic, or actionable, is a known challenge in online social spaces (Sheth et al., 2022) and in games specifically (Beres et al, 2021; Kocielink, et al., 2024). For example, there are some scholars in the field who have suggested that trash talking should not be considered toxicity at all, but rather an accepted part of certain kinds of gaming experiences (Deloy, Nino, & San Isidro, 2022; Fennell, 2021). Similarly, it is unclear if more severe offenses like sexual harassment are seen simply as "normalized" experiences within these spaces (so-called "gaming banter") or are considered by the industry as a harm that necessitates consistent

and decisive punitive action (Beres et al, 2021; Kowert, Kilmer, & Newhouse, in press). It is vitally important that player and professional groups share consensus as to what is, and what is not, considered acceptable within their community, not only to support the perceived fairness and legitimacy of any potential punitive action but to strengthen compliance with the rules and code of conduct set in place for any particular environment (Aguerri et al., 2023).

This potential misalignment on harms between the service provider (i.e., company) and user has seen some attention in social media spaces (Aguerri et al., 2023), but there remains limited research on whether industry professionals and players are aligned in the severity of different types of harms in gaming spaces. To gain insight into these lingering questions and potential discrepancies, we conducted an online survey evaluating the perception of severity and prevalence of a range of online harms among game players and trust and safety professionals from the gaming industry.

03

Current study

An online survey was created to record perceptions of harm prevalence and severity in online gaming spaces for player communities and game industry professionals who specialize in the field who specialize in player safety (e.g., trust and safety professionals).

Participants

Game players were recruited via Prolific, an online research platform that facilitates participant recruitment for social scientists. Integrating with Qualtrics, we recruited a representative sample of adults (n = 100), predominantly located in the UK (88%) and United States (9%). The mean age of players was 33.98 (SD = 8.69, 18-60 years old). Regarding gender, 64% identified themselves as male (N = 64), 34% as female (N = 34), and 2% (N = 2) as nonbinary or another gender). When asked about the kinds of multiplayer spaces they predominantly engaged in, the majority of players (64%) reported console or desktop AAA titles, followed by console or desktop indie titles (21%), mobile games (14%), and game-adjacent platforms (1%).

To recruit trust and safety (T&S) professionals in the gaming industry (n = 34), we reached out via professional channels, such as professional Discord servers, newsletters, and LinkedIn. The mean age of T&S professionals was 35.27 (SD = 9.25, 19-57 years old). Most T&S participants were male (50%, N = 17), followed by female (38%, N = 13), nonbinary or other gender (6%, N = 2), and preferred not to say (6%, N = 2).

Methods and Measures

To more effectively parse the data, harms were split into two categories: gameplay harms and social harms. These categories are not wholly discrete, rather in this research they serve as general guidelines through which observations can be made. Gameplay harms are behaviors directly related to gameplay or gameplay experiences and are meant to interfere with the way in which people are playing the game. These actions happen primarily within the game space itself. Gameplay harms include aiding the enemy, behavioral spamming, cheating, contrary play, flaming, griefing, inappropriate role playing, inhibiting team, trash talking, and verbal spamming.

Social harms are behaviors that are primarily intended to harm the player as an individual rather than impede gameplay. These harms can happen within the game play space itself or on a third party site or platform (e.g., Discord). Social harms include child sexual abuse material (CSAM), coordinated inauthentic behavior (CIB), doxxing, fraud, gender based violence, hate raiding, hate speech, incitement of violence, impersonation, misinformation, sexual harassment, swatting, and threats of violence.

A full list of categories and definitions can be found in Table 1.

It is important to note that these groups of harms are not mutually exclusive, but rather categorical distinctions to help us understand trends around online harms in and around online gaming spaces.

Participants were asked to rate prevalence and severity of each category of harms in online gaming spaces. For players, prevalence was defined as the rate at which they encounter the action in gaming spaces. For trust and safety professionals, prevalence was defined as the rate at which they encountered the actions in their work, rather than personal experience. For all participants, severity was defined as the extent to which they believe their action is harmful to the well-being (physical and psychological) of the target, witnesses, and wider community.

Table 1. Category of gameplay and social harms and definitions.

Gameplay Harms	
Aiding the enemy	Strategically aiding the opposing team
Behavioral Spamming	Using the same in-game move, often to the consternation of others
Cheating	Using methods to create an advantage beyond normal gameplay in order to make the game easier for oneself
Contrary play	Playing outside of what is intended by most players
Flaming	Presenting emotionally fueled or contrary statements
Griefing	Irritating and/or harassing other players by using the game in unintended ways

Inappropriate role playing	Pretending to be a different person to obtain a specific reaction or not abiding by the norms of the community
Inhibiting team	Inhibiting your team from being successful in winning
Trash talking	Putting down or making fun of other players
Verbal spamming	Sending the same verbal message or using the same in-game move
Social Harms	
Child sexual abuse material (CSAM)	Imagery or videos which show a person who is a child engaged in or is depicted as being engaged in explicit sexual activity
Coordinated inauthentic behavior (CIB)	A manipulative communication tactic that uses a mix of authentic, fake, and duplicated social media accounts to operate as an adversarial network across multiple social media platforms
Doxxing	Publicly sharing and/or publishing another player's identifying information
Fraud	Intentional deception to secure unfair or unlawful gain or to deprive a victim of a legal right.
Gender based violence	Violence directed against a person because of that person's gender or violence that affects a person of a particular gender disproportionately
Hate raiding	Purposefully infiltrating the game space of another with the intention of spreading hate or harassment
Hate speech	Insults based on religion, ethnicity, nationality, or other personal information

Incitement of violence	Speech, words, or behaviors that encourage the immediate risk of harm to another person
Impersonation	The act of pretending to be another person for the purpose of entertainment or fraud
Misinformation	Repeatedly sharing game-unrelated chat
Sexual harassment	Insults or comments based on gender, including threats, criticism, or stalking
Swatting	Calling emergency services in an attempt to dispatch armed police officers to a particular address
Threats of violence	Threats of physical abuse, vandalism, possession or use of weapons, or other dangerous action

Results

Prevalence of Harms

For gameplay harms, significant differences were found for trash talking, inappropriate role playing, contrary play, and flaming (see Figure 1), with industry professionals rating these items as more prevalent. Effect sizes were calculated to estimate the practical significance of these differences - effect sizes were medium (Cohen's d = .51 - .74; see Table 2), suggesting the presence of real-world relevance of these differences.

Differences in the perception of prevalence between players and industry professionals were also found for social harms, with professionals reporting higher prevalence rates for CIB, CSAM, doxxing, fraud, gender based violence, hate raiding, hate speech, impersonation, incitement of violence, sexual harassment, and swatting (see Figure 2). Effect sizes ranged from medium to large (d = .52 - 1.01; see Table 3), again suggesting practical differences in the two group's perceptions of prevalence.

Figure 1. Differences between T&S and player perceptions of behavior frequency of gameplay harms

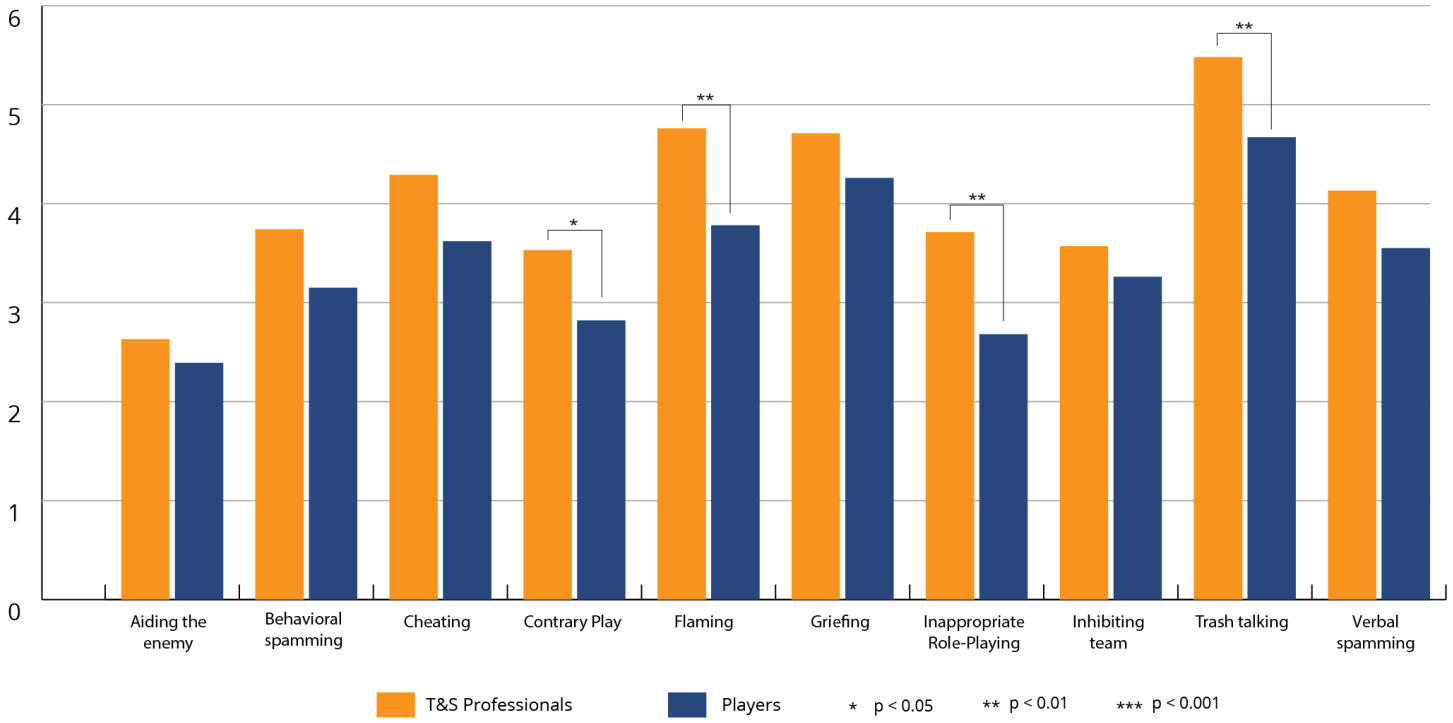
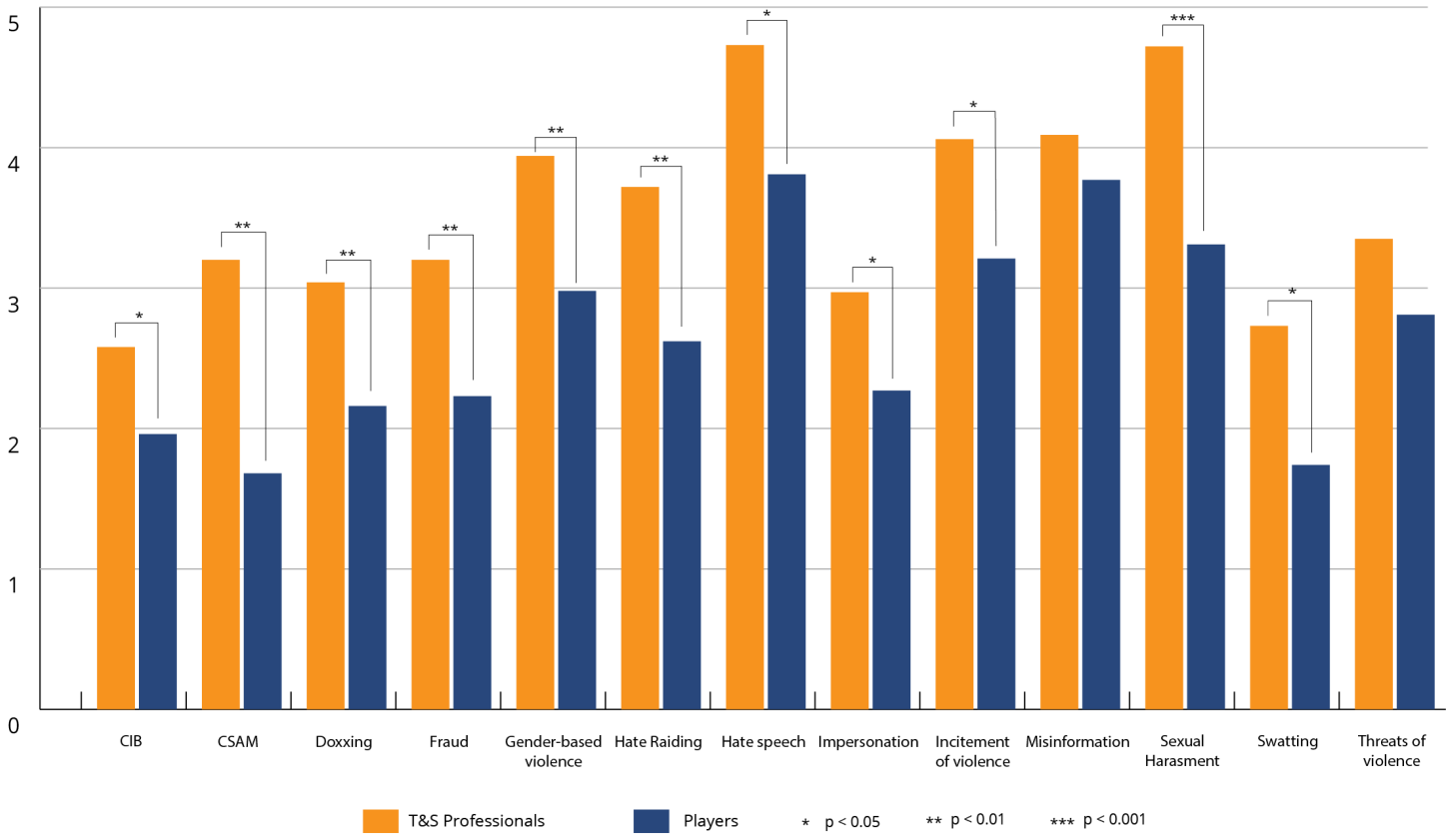


Figure 2. Differences between T&S and player perceptions of behavior frequency of social harms



Severity of Harms

For gameplay harms, significant differences were found for inappropriate role-playing, contrary play, behavioral spamming, and cheating (see Figure 3), with industry professionals rating these items as more prevalent. The effect sizes for these four types ranged from medium to large (d = .51 - .92; see Table 4), which indicates there are likely practically meaningful differences between player and T&S perception of harm severity in these areas.

Differences in the perception of severity between players and industry professionals were also found for social harms, with professionals reporting higher severity rates for behavioral spamming, cheating, CSAM, doxxing, gender based violence, hate raiding, hate speech, incitement of violence, misinformation, and swatting (see Figure 4). The effect sizes for these differences ranged from small to medium (d = .32 - .68; see Table 5).

Figure 3. Differences between T&S and player perceptions of behavior severity for gameplay harms

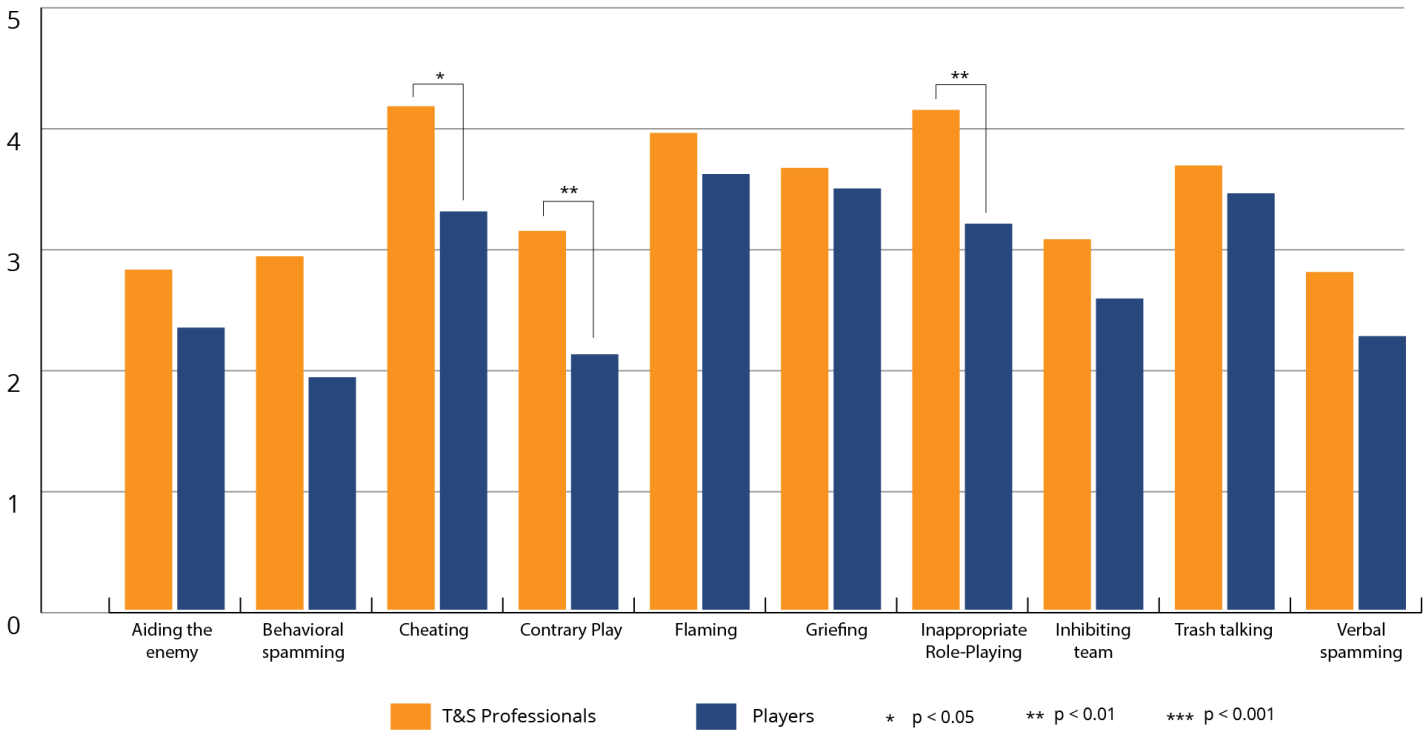
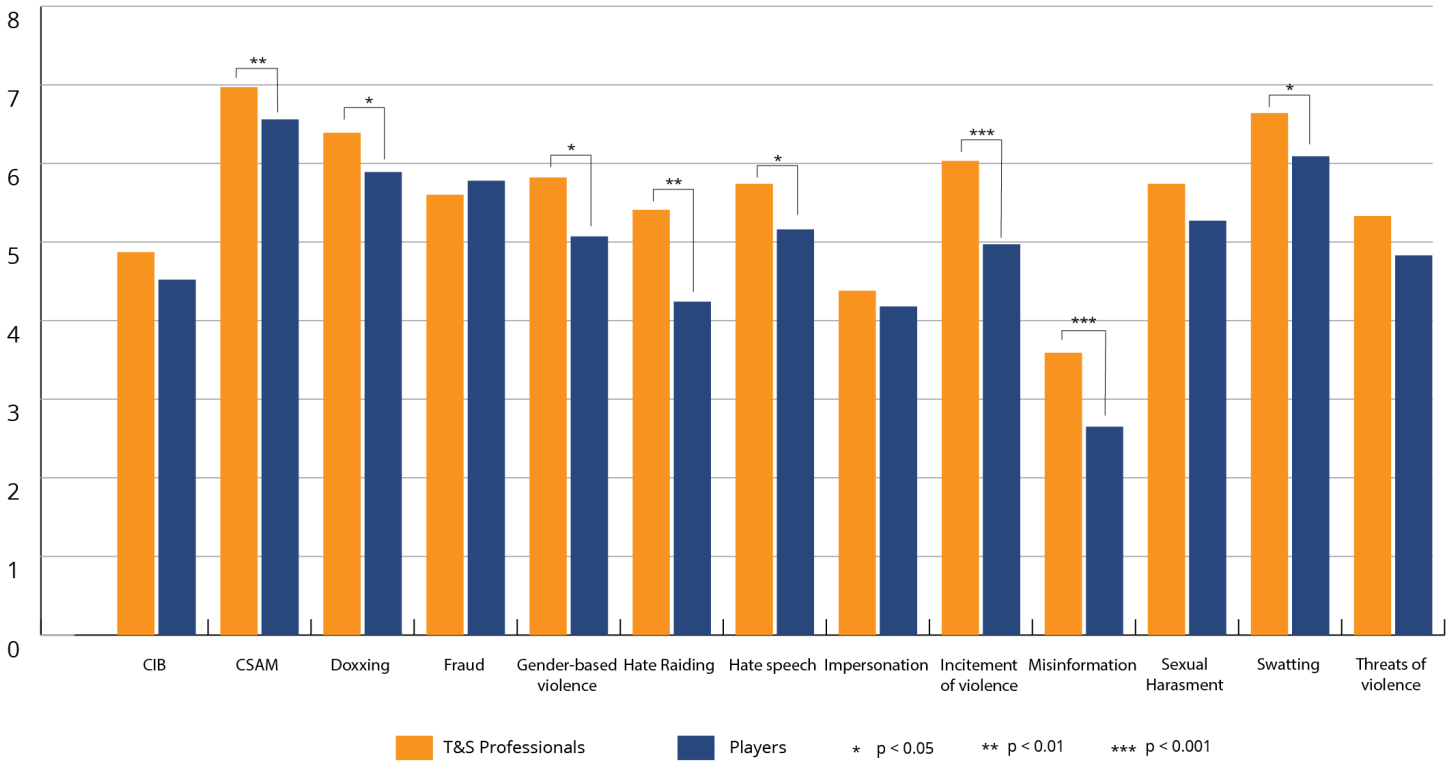


Figure 4. Differences between T&S and player perceptions of behavior severity for social harms



Discussion

It has been questioned whether game players, and those responsible for their safety, are aligned in their beliefs about what behavior in these spaces is truly harmful and deserving of action on behalf of the gaming platforms. In this work, we assessed the perception of prevalence and severity of a range of harms in order to gain insight into these potential disparities.

Industry professionals reported higher prevalence and severity rates than players. Industry professionals consistently reported higher prevalence rates than players for gameplay and social harms, with social harms being rated as more severe offenses than gameplay harms.

Disparities in the perception of the prevalence of harms between groups suggests that trust and safety professionals in the gaming industry have been effective at shielding players from some of the most problematic harm categories, including sexual harassment, CSAM, and swatting. It is important to note that the effect sizes for these differences in perceived prevalence and severity ranged from medium to large, while the significant differences in social harms severity were small to medium. This means that the practical differences in ratings across player and professional groups were smaller for the severity of social harms, suggesting that players and professionals may be more aligned on perceived seriousness of social harms than they are those of gameplay harms.

There are several takeaways from this work that can help inform T&S teams and the broader industry to support effective mitigation strategies for harmful content and cultivate thriving gaming communities:

- **Player and Professional groups agree on the severity of in-game harms.** For most harm categories, player and professional groups were aligned in their perception of their severity, with few exceptions. It is important to note that while industry professionals rated many categories of social harm as more severe, effect sizes indicated players and professionals are actually relatively aligned

- **Trust and safety teams are effectively moderating some forms of harm in gaming spaces.** A greater perceived prevalence of a range of gameplay and social harms from trust and safety teams than player groups suggests they are seeing, and shielding, players from a significant number of harms across categories.

- **Trust and safety teams are the most effective at moderating the most severe harms.** Industry professionals report seeing higher prevalence rates for the most severe social offenses, including CSAM, doxxing, hate speech, incitement of violence and swatting, suggesting they are effectively shielding player groups from a significant number of these offenses.

Limitations and Future Directions

While this work sheds light on many unanswered questions around trust and safety concerns in gaming spaces, there are several limitations to acknowledge. The small sample size for industry professionals who work in trust and safety (N = 34) limits the generalizability of these findings. Furthermore, although a brief definition of each type of harm was provided for participants, participants may have a different frame of reference or understanding about what constitutes each category of harm and this could have impacted their ratings of perception and severity. Lastly, we only assessed participants' perception of harms in the online gaming spaces. While this was done in order to generate a broad landscape analysis, future work looking more in-depth into specific communities would help clarify variation across gaming and game-adjacent spaces. This would be particularly valuable as we know that player behavior can change depending on community guidelines in online spaces (Kowert, Botelho, & Newhouse, 2022; Smith et al., 2021).

It is also worth noting that while we hypothesize differences in perception between player and industry groups indicate effectiveness in moderation, it is also possible that because trust and safety professionals are so regularly exposed to such problematic content that it inflates such professionals' understanding of the prevalence rates. Additional research would be needed to

clarify if the disparities we found here are due to effectiveness of their efforts or over-estimation.

Conclusion

A greater understanding of the areas of alignment (and misalignment) of trust and safety professionals in the gaming industry and the players in their communities is vital to the continued development of effective policies and strategies to keep gaming communities safe and thriving. When these groups are not aligned on safety-related concerns, players may be less likely to follow community guidelines themselves, or file actionable reports when they see other players violate guidelines. This work suggests that trust and safety professionals take harms within their spaces very seriously, particularly social harms. Disparities in frequency indicate that they are also effective at protecting players from encountering many of these harms in the first place. Both of these results are encouraging; however, the gaming industry is still struggling with moderation at scale given the prevalence of harms still reported by player groups.

References

Aguerri, J. C., Miró-Llinares, F., & Gómez-Bellvís, A. B. (2023). Consensus on community guidelines: an experimental study on the legitimacy of content removal in social media. Humanities and Social Sciences Communications, 10(1). https://doi.org/10.1057/s41599-023-01917-2

Anti-Defamation League. (2024). Hate Is No Game: Hate and Harassment in Online Games 2023. https://www.adl.org/resources/report/hate-no-game-hate-and-harassment-online- games-2023

Beres, N. A., Frommel, J., Reid, E., Mandryk, R. L., and Klarkowski, M. (2021). "Don't you know that you're toxic: Normalization of toxicity in online gaming," in Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems 1–15. doi: 10.1145/3411764.3445157

Deloy, E. D., Nino, S., & San Isidro, D. D. N. P. (2022). The culture of trash talks among Dota players: An ethnography. International Journal of Research Publications, 109.

Fennell, M. (2021). Trash talk or smack talk: The language of competitive sport. Journal of Higher Education Athletics & Innovation, 1(9), 33-48.

Kocielnik, R., Li, Z., Kann, C., Sambrano, D., Morrier, J., Linegar, M., Taylor, C., Kim, M., Naqvie, N., Soltani, F., Dehpanah, A., Cahill, G., Anandkumar, A., & Alvarez, R. M. (2024). Challenges in moderating disruptive player behavior in online competitive action games. Frontiers in Computer Science, 6. https://doi.org/10.3389/fcomp.2024.1283735

Kowert, R., Botelho, A., & Newhouse, A. (2022). Breaking the building blocks of hate: A case study of Minecraft servers. A report from the Anti-Defamation League.(ADL) center of Technology and Society. https://www.adl.org/resources/report/breaking-building-blocks-hate- case-study-minecraft-servers

Kowert, R., Kilmer, E., & Newhouse, A. (2024). Taking it to the extreme: prevalence and nature of extremist sentiment in games. Frontiers in Psychology, 15. https://doi.org/10.3389/fpsyg.2024.1410620

Kowert, R., Kilmer, E., & Newhouse, A. (2024, January 3). Culturally justified hate: Prevalence and mental health impact of dark participation in games. Proceedings of the 57th Hawaii International Conference on System Sciences. https://hdl.handle.net/10125/106708

Sheth, A., Shalin, V. L., and Kursuncu, U. (2022). Defining and detecting toxicity on social media: context and knowledge are key. Neurocomputing 490, 312–318. doi: 10.1016/j.neucom.2021.11.095

Smith, J., Krasodomski-Jones, A., Olanipekun, M., & Judson, E. (2021). A picture of health: Measuring the comparative health of online spaces. https://demos.co.uk/research/a-picture-of-health-measuring-the-comparative-health-of-online-spaces/

Suler, J. (2004). The Online Disinhibition Effect. CyberPsychology & Behavior, 7(3), 321–326. https://doi.org/10.1089/1094931041291295

Unity. (2023). Toxicity In Multiplayer Games Report. https://create.unity.com/toxicity-report?utm_source=substack&utm_medium=email