A Hierarchy of Responsibility for Harm Mitigation in Design of Video Games

Zinan Zhang

College of Information Sciences and Technology, The Pennsylvania State University, zzinan@psu.edu

1 POSITION STATEMENT

Video games are often celebrated for providing entertainment, creativity, and social connection. However, they also produce a wide range of harms, including exposure to violent and sexualized content [1, 2], manipulation through dark design patterns [10, 11], and persistent toxicity and harassment between players [6, 8]. These harms are often treated as isolated incidents or as the result of individual bad actors. In contrast, my research approaches harm as a systemic outcome of sociotechnical design, shaped by the interaction between game mechanics, participatory creation practices, platform governance, and policy environments. I position harm mitigation—rather than post-hoc moderation or punishment—as a central responsibility of responsible design.

Across my work, I examine how harm emerges in user-generated game ecosystems and how it might be mitigated through design decisions made at multiple levels: platform architecture, designer practices, and player interaction. I argue that effective harm mitigation requires shifting responsibility upstream—from reacting to harm after it occurs toward proactively shaping design environments that reduce the likelihood, scale, and severity of harm.

A central strand of my work investigates harmful design in user-generated game platforms, particularly Roblox, a popular child-oriented ecosystem where anyone can design, develop, and publish games (e.g., [7, 9, 12, 14]). While this participatory model enables creativity and learning, it also allows harmful ideas to emerge, evolve, and spread through community co-ideation. In my prior research, I found that seemingly benign design ideas can escalate into harmful ones through collaborative ideation—for example, a generic building model evolving into simulations of real-world traumatic events such as the Twin Towers or 9/11 [13]. Within creator communities, designers not only co-develop these ideas but also share strategies for bypassing moderation systems to ensure their publication. In these cases, responsibility does not lie with a single actor. Instead, it is distributed across: primary designers, who initially implement harmful or risky mechanics; co-designers, whose ideation practices may amplify and normalize harm; and platforms, which often focus moderation on published content while leaving the underlying design mental models and ideation cultures largely unregulated. This work highlights that responsible design must extend beyond assessing a creator's intent. Platforms must also consider how their tools, norms, and governance structures shape what designers imagine as acceptable or desirable design.

Another line of my research focuses on unintended consequences of game design, particularly in relation to toxicity and harassment. Harmful behavior in games is often framed as the result of individual bad actors [4, 5], yet my findings show that design choices frequently create the conditions, triggers, and affordances that enable such harm [15]. For example, competitive mechanics that require heavy interdependence between players can increase blame and hostility when outcomes are negative. Communication features such as voice chat can expose players' gender, making female players particularly vulnerable to harassment and targeting. Seemingly minor interaction affordances—such as avatar gestures, repeated pinging systems, or spatial mechanics like "teabagging"—can normalize or intensify toxic behavior. These cases illustrate that ethical responsibility cannot be limited to designers' original intentions. Design features may be repurposed or misused in harmful ways, and responsibility must therefore also account for how designs are taken up, interpreted, and enacted by players at scale. This shifts responsibility beyond individual designers to include platform-level decisions about affordances, defaults, and social norms.

Across both research threads, a common pattern emerges: harm mitigation is often deferred to moderation systems that intervene after harm has already occurred. Content is removed, accounts are banned, or players are reported—but the underlying design conditions remain unchanged. This reactive approach places a disproportionate burden on players, particularly marginalized or younger players, who must endure harm before any corrective action is taken. Prior scholarship has described this limitation and called for safety by design [3], emphasizing that systems should be designed to minimize harm by default rather than relying on enforcement alone. Building on this work, I argue that harm mitigation must be understood as a design responsibility distributed across the sociotechnical stack, rather than a function delegated solely to moderators or automated filters. To make sense of these dynamics, I conceptualize harm mitigation as operating within a hierarchy of cascading responsibility [16]:

Policy level. At the top of the hierarchy, public policy and regulatory frameworks shape how platforms define safety, accountability, and acceptable risk. Policies influence platform incentives, enforcement priorities, and the resources allocated to harm mitigation.

Platform and service provider level. Platforms translate policy into governance structures, moderation systems, designer tools, and economic incentives. Platform-level design decisions shape what kinds of games are easy to make, what kinds of behavior are rewarded, and how responsibility is distributed between designers and players.

Designer level. Game designers operate within these constraints, making concrete decisions about mechanics, interaction affordances, monetization strategies, and social systems. Their choices directly shape players' experiences and exposure to harm.

Player level. Players enact, interpret, and sometimes misuse designs. When harm mitigation fails at higher levels, players—especially those with less power or fewer resources—bear the consequences.

When harm mitigation fails upstream, responsibility cascades downward. Players are asked to report, block, tolerate, or adapt, even when harms could have been mitigated through better design or governance earlier in the process. My research argues that responsible design requires reversing this cascade, placing greater responsibility on upstream actors to prevent harm before it materializes.

My work contributes to responsible design and harm mitigation research by empirically demonstrating how harm emerges through participatory design, unintended consequences, and platform governance gaps in user-generated game ecosystems. By foregrounding harm mitigation rather than punishment, my research highlights opportunities to intervene earlier in the design lifecycle and to distribute responsibility more equitably across stakeholders. I am particularly interested in engaging with this workshop to discuss how harm mitigation principles can be operationalized in responsible design—across policy, platform governance, designer tooling, and player-facing design. I see this workshop as an opportunity to collaboratively explore frameworks, methods, and design strategies that move beyond reactive moderation toward proactive, responsible design that meaningfully reduces harm.

REFERENCES

[1] Christopher P. Barlett, Richard J. Harris, and Callie Bruey. 2008. The effect of the amount of blood in a violent video game on aggression, hostility, and arousal. *Journal of Experimental Social Psychology* 44, 3 (May 2008), 539–546. https://doi.org/10.1016/j.jesp.2007.10.003

[2] Jonathan Burnay, Brad J. Bushman, and Frank Larøi. 2019. Effects of sexualized video games on online sexual harassment. *Aggressive Behavior* 45, 2 (March 2019), 214–223. https://doi.org/10.1002/ab.21811

[3] Andrew Hale, Barry Kirwan, and Urban Kjellén. 2007. Safe by design: where are we now? *Safety science* 45, 1–2 (2007), 305–327. https://doi.org/10.1016/j.ssci.2006.08.007

[4] Bastian Kordyaka, Samuli Laato, Juho Hamari, Tobias Scholz, and Björn Niehaves. 2023. What drives gamer toxicity? Essays from players. (April 2023). Retrieved November 30, 2023 from https://trepo.tuni.fi/handle/10024/151587

[5] Bastian Kordyaka, Samuli Laato, Sebastian Weber, and Bjoern Niehaves. 2023. What constitutes victims of toxicity - identifying drivers of toxic victimhood in multiplayer online battle arena games. *Frontiers in Psychology* 14, (2023). Retrieved November 30, 2023 from https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1193172

[6] Yubo Kou. 2020. Toxic Behaviors in Team-Based Competitive Gaming: The Case of League of Legends. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, November 02, 2020. ACM, Virtual Event Canada, 81–92. https://doi.org/10.1145/3410404.3414243

[7] Yubo Kou, Yingfan Zhou, Zinan Zhang, and Xinning Gui. 2024. The Ecology of Harmful Design: Risk and Safety of Game Making on a Metaverse Platform. In *Designing Interactive Systems Conference*, July 2024. ACM, IT University of Copenhagen Denmark, 1842–1856. https://doi.org/10.1145/3643834.3660678

[8] Elena Koung, Zinan Zhang, Xinning Gui, and Yubo Kou. 2025. Gendered Toxicity in Competitive Gaming: Women's Perceptions and Responses. *Proc. ACM Hum.-Comput. Interact.* 9, 6 (October 2025), GAMES015:415-GAMES015:446. https://doi.org/10.1145/3748610

[9] Qiurong Song, Zinan Zhang, Rie Helene Lindy Hernandez, Xinning Gui, and Yubo Kou. 2025. How Predatory Monetization Designs Manifest in Child-Friendly Video Games. 2025. Twenty-First Symposium on Usable Privacy and Security (SOUPS). https://www.usenix.org/conference/soups2025/presentation/song

[10] José P. Zagal, Staffan Björk, and Chris Lewis. 2013. Dark patterns in the design of games. *Foundations of Digital Games* (2013).

[11] Gloria Xiaodan Zhang, Yijia Wang, Taro Leo Nakajima, and Katie Seaborn. 2025. First Contact with Dark Patterns and Deceptive Designs in Chinese and Japanese Free-to-Play Mobile Games. *Proc. ACM Hum.-Comput. Interact.* 9, 6 (October 2025), GAMES025:730-GAMES025:755. https://doi.org/10.1145/3748620

[12] Zinan Zhang, Xinning Gui, Junnan Yu, Sunhye Bai, and Yubo Kou. 2025. Dangerous Playgrounds: Child Players' Encounters with Design-Mediated Risks on User Generated Game Platforms and Their Safety Practices. *Proceedings of the 25th ACM Conference on Interaction Design and Children* (2025). https://doi.org/10.1145/3713043.3728858

[13] Zinan Zhang, Sam Moradzadeh, Xinning Gui, and Yubo Kou. 2024. Harmful Design in User-Generated Games and its Ethical and Governance Challenges: An Investigation of Design Co-Ideation of Game Creators on Roblox. *Proc. ACM Hum.-Comput. Interact.* 8, CHI PLAY (October 2024), 1–31. https://doi.org/10.1145/3677076

[14] Zinan Zhang, Sam Moradzadeh, Xinning Gui, and Yubo Kou. 2025. More Than Just Microtransactions: Predatory Monetization in User-Generated Games. *Proc. ACM Hum.-Comput. Interact* 9, (October 2025). https://doi.org/10.1145/3748626

[15] Zinan Zhang, Sam Moradzadeh, Andrew Woan, and Yubo Kou. 2024. Toxicity by Game Design: How Players Perceive the Influence of Game Design on Toxicity. *Proc. ACM Hum.-Comput. Interact.* 8, CHI PLAY (October 2024), 345:1-345:31. https://doi.org/10.1145/3677110

[16] Zinan Zhang, Qiurong Song, Rie Helene Lindy Hernandez, Yunhan Liu, Elena Koung, Junnan Yu, Sunhye Bai, Yubo Kou, and Xinning Gui. 2026. Player Safety by Design: Co-Designing Child-Centered Safety Mechanisms with Children. (2026). Retrieved March 12, 2026 from https://www.researchgate.net/profile/Zinan-Zhang-12/publication/401586155_Player_Safety_by_Design_Co-Designing_Child-Centered_Safety_Mechanisms_with_Children/links/69a9c23984431b5258b7cc55/Player-Safety-by-Design-Co-Designing-Child-Centered-Safety-Mechanisms-with-Children.pdf