Literature Review Summary Table

As discussed in our commentary, we conducted a pilot test literature review, using the History of Trust and Safety Archive. Our review focused on harassment and cyberbullying literature published in Communications and Human Computer Interaction venues. This table includes reference information, a summary of the findings, and other details for the 75 peer-reviewed documents we reviewed in our search. All these items can be found in the History of Trust & Safety Zotero archive.

Citation	Disciplinary Subfolder(s)	Population and/or Content Studied	Method	Findings Summary
Aguerri, J. C., Santisteban, M., & Miró-Llinares, F. (2023). The Enemy Hates Best? Toxicity in League of Legends and Its Content Moderation Implications. European Journal on Criminal Policy and Research. https://doi.org/10.1007/s10610-023-09541-1	Communications	Matches from a European League of Legends official tournament	Observational	Disruptive behavior (evidenced in 70% of matches) is common in League of Legends but harmful behavior such as hate speech is considerably less common (10.9%). Moderation interventions should consider the distinction between more and less severe disruptive behavior and work to better understand what forms of harassment most impact users.
Aliapoulios, M., Take, K., Ramakrishna, P., Borkan, D., Goldberg, B., Sorensen, J., Turner, A., Greenstadt, R., Lauinger, T., & McCoy, D. (2021). A large-scale characterization of online incitements to harassment across platforms. Proceedings of the 21st ACM Internet Measurement Conference, 621–638. https://doi.org/10.1145/3487552.3487852	Human Computer Interaction (HCI)	14,679 user-generated incitements to harassment from Boards, Discord, Gab, Pastes, Telegram	Observational	Over 50% of incitements by coordinated attackers include a call to report the target to legal authorities or the social media platform. The strategies and approaches attackers use to target others are dynamic and authorities and researchers must consider how anti-harassment policies may be ineffective or actively abused by malicious actors.
Bhandari, A., Ozanne, M., Bazarova, N. N., & DiFranzo, D. (2021). Do You Care Who Flagged This Post? Effects of Moderator Visibility on Bystander Behavior. Journal of Computer-Mediated Communication, 26(5), 284–300. https://doi.org/10.1093/jcmc/zmab007	Human Computer Interaction (HCI) and Communications	582 MTurk participants	Experiment	The experiments found that telling users who or what moderated a comment, AI or human, made them less likely to flag future unmoderated content, suggesting that in some cirumstances transparency in moderation activities may inhibit bystander intervention.
Blackwell, L., Dimond, J., Schoenebeck, S., & Lampe, C. (2017). Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. Proceedings of the ACM on Human-Computer Interaction, 1(CSCW), 1–19. https://doi.org/10.1145/3134659	Human Computer Interaction (HCI)	25 virtual reality (VR) (Oculus) users living in the USA	Interview	How users define experiences of online harassment, and lived impacts of harassment, vary widely and are highly personal, particularly in the context of VR where social norms are developing and reporting/behavioral enforcement mechansisms are limited.
Blackwell, L., Ellison, N., Elliott-Deflo, N., & Schwartz, R. (2019). Harassment in Social Virtual Reality: Challenges for Platform Governance. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 100:1-100:25. https://doi.org/10.1145/3359202	Human Computer Interaction (HCI)	18 HeartMob users	Interview	Testing the HeartMob platform revealed how labelling experiences of harassment can legitimize dominant experiences of harassment, but also exclude or alienate experiences the system does not recognize as harassment. Labelling also helps bystanders recognize harassment and set expectations for interaction in the community.
Casula, P., Anupam, A., & Parvin, N. (2021). "We found no violation!": Twitter's Violent Threats Policy and Toxicity in Online Discourse. Proceedings of the 10th International Conference on Communities & Technologies - Wicked Problems in the Age of Tech, 151–159. https://doi.org/10.1145/3461564.3461589	Human Computer Interaction (HCI)	The study was primarily theoretical/conceptual, but utilized a small case study of 3 tweets which the authors argue demonstrate how the Twitter Violent Threats Policy overlooks harms to women and other marginalized groups	Case Study and Theory/Conceptual	Implementation of the Twitter Violent Threats Policy in its current form fosters a toxic environment of pre-emptive self-censorship, despite free speech claims, demonstrated by continued harassment against political figures and the general figures. The authors argue that the platform needs to consider the rhetorical situation content is presented in in order to implement
Chadha, K., Steiner, L., Vitak, J., & Ashktorab, Z. (2020). Women's Responses to Online Harassment. International Journal of Communication, 14(0), Article 0.	Human Computer Interaction (HCI) and Communications	23 women university students who identified having experienced harassment/cyberbullying	Interview	Women deploy numerous strategies in anticipation of online harassment including self-censorship and withdrawing from participation, urging more attention to fostering women's safe and equitable participation in online spaces.

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., & Gilbert, E. (2017). You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. Proceedings of the ACM on Human-Computer Interaction, 1(CSCW), 1–22. https://doi.org/10.1145/3134666	Human Computer Interaction (HCI)	100 million posts scraped from r/fatpeoplehate and r/CoonTown generated January - December 2015	Observational	After the 2015 Reddit-backed shutdown of r/CoonTown and r/fatpeoplehate, the study found that communities who adopted migrants from these communities did not see changes in the amounts or types of hate speech produced in their communities. suggesting that carefully considered bans can be effective in preventing hate speech.
Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017). Mean Birds: Detecting Aggression and Bullying on Twitter (arXiv:1702.06877). arXiv. http://arxiv.org/abs/1702.06877	Human Computer Interaction (HCI)	Two Twitter tweet corpuses (one of 1 million random treats, one containing 650,000 tweets containing language from hate speech glossary) generated June - August 2016	Observational	Bullying users post let, participate in fewer communities overall, and have less social connections than non-bullying counterparts. The classification alogorithms tested in this inquiry can identify bullying and harassing users with a high degree of accuracy (90% or higher).
Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, 71–80. https://doi.org/10.1109/SocialCom-PASSAT.2012.55	Human Computer Interaction (HCI)	Developed and tested Lexical Syntactic Feature (LSF) architecture for identifying undesirable content and potentially offensive users utilizing comments from 2,175,474 unique users comments from top 18 most viewed YouTube videos.	Observational and Conceptual/Theoretical	The Lexical Syntactic Feature (LSF) architecture developed and tested in this inquiry improved sentence and user level offensive content detection as compared to other existing methods. Utilizing this technique on social media could improve the efficiency and accuracy of current content moderation approaches.
DeCook, J. R. (2022). R/WatchRedditDie and the politics of reddit's bans and quarantines. Internet Histories, 6(1–2), 206–222. https://doi.org/10.1080/24701475.2021.1997179	Human Computer Interaction (HCI) and Communications	3 year ethnographic study (2015-2018) of policy change and community response in r/WatchRedditDie community,	Observational	In response to platform changes, r/WatchRedditDie members developed a sense of community around a shared understanding of free expression and antagonism towards the platform itself. This community is evidence of the challenges shared governance between platforms and users can face.
Dinakar, K., Jones, B., Havasi, C., Lieberman, H., & Picard, R. (2012). Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying. ACM Transactions on Interactive Intelligent Systems, 2(3), 1–30. https://doi.org/10.1145/2362394.2362400	Human Computer Interaction (HCI)	Scraped YouTube comments on ideologically controversial videos, had a middle school USA based population code the content for cyberbullying material, and developed new language processing model informed by the unique characteristics of cyberbullying and proposed BullySpace resource platform.	Observational, Survey, and Design	This inquiry proposed BullySpace, a model developed in consideration of the unique contextual and content considerations of cyberbullying as compared to other forms of online harassment. Based upon deploying the model on a social media sample and user interviews, the authors propose tailored, realtime support and communication with victims as a valuable approach for improving user
Duguay, S., Burgess, J., & Suzor, N. (2020). Queer women's experiences of patchwork platform governance on Tinder, Instagram, and Vine. Convergence, 26(2), 237–252. https://doi.org/10.1177/1354856518781530	Human Computer Interaction (HCI)	Interviews with 20 self identifed queer women users on Tinder, Instagram, and Vine along with observing and analyzing platform policies and architectures	Interview and Conceptual/Theoretical	To address how queer women experience harassment and violence on online platforms, social media platforms should critically re-assess the criteria they use to identify harassment, response policies, communication with victims, and other forms of platform architectures to address bad behavior and support more inclusive online cultures.
DuVal Smith, A. (1999). Problems of conflict management in virtual communities. In Problems of conflict management in virtual communities (1st ed., pp. 135–166). Routledge. https://www-taylorfrancis-com.proxy2.cl.msu.edu/chapters/edit/10.4324/9780203194959-16/problems-conflict-management-virtual-communities-anna-duval-smith	Communications	1993-1998 participant observation on MicroMUSE, IRC, and MUD forums.	Interview, Observational, and Conceptual/Theoretical	Managing the conflict that accompanies human diversity is especially important to address in online communities because online communities are especially likely to be highly diverse. Online communities may benefit from deploying real world techniques for conflict management including incorporating third party liasions to weigh in on community conflits.
Edwards, L., & Leatherman, A. (2009). ChatCoder: Toward the Tracking and Categorization of Internet Predators. https://www.semanticscholar.org/paper/ChatCoder%3A-Toward-the-Tracking-and-Categorization-Edwards-Leatherman/01ce28a8f3fbd949f061e0ea5ccc92432a8d1f9d	Human Computer Interaction (HCI)	Coded online chat room conversations to identify features of user-generated text which may predict sexual predation. Used this study to experimentally test a software tool to identify predatory and non-predatory content along with predicting individual users' likelihood to generate predatory content.	Observational, Experimental and Conceptual/Theoretical	Deployed a software to detect predator and victim conversations. The tool was most effective in distinguishing between victim and predator discussion text. The authors also identified four distinct archetypes of online predation using this software.

Felmlee, D., DellaPosta, D., Inara Rodis, P. d. C., & Matthews, S. A. (2020). Can Social Media Anti-abuse Policies Work? A Quasi-experimental Study of Online Sexist and Racist Slurs. Socius, 6. https://doi.org/10.1177/2378023120948711	Communications	3.6 million Tweets produced one month before through one month after (November 17 2017 to January 18 2018) Twitter rule change regarding abusive content	Observational and Experimental	The study found that negative sentiments surrounding African Americans decreased after the new deployment of this policy, suggesting that policy revisions alongside changing social norms and social legitimacy surrounding undesirable content can serve as a pathway to prevention and mitigation.
Feng, C., & Wohn, D. Y. (2020). Categorizing Online Harassment Interventions. 2020 IEEE International Symposium on Technology and Society (ISTAS), 255–265. https://doi.org/10.1109/ISTAS50296.2020.9462206	Human Computer Interaction (HCI)	Systematic literature review of online harassment interventions consisting of 17 peer-reviewed manuscripts which empiricallty examined 17 investigations into bystander intervention in the context of online harassment	Theoretical/Conceptual	Based on peer-reviewed scholarship's current focus on negative outcomes related to interventions, and publicly accessible datasets, the authors suggest platforms should incorporate modelling good behavior for users, prompting reflection before posting flagged material, and anonymous content reporting.
Finn, J. (2004). A Survey of Online Harassment at a University Campus. Journal of Interpersonal Violence, 19(4), 468–483. https://doi.org/10.1177/0886260503262083	Communications	339 University of New Hampshire undergraduate students	Survey	Sexual minority students experience higher rates of stranger-based sexual harassment, prompting further research on differences in online harassment experiences and outcomes in LGBTQ+ and non-LGBTQ+ populations and more education for students about online violence and reporting pathways.
Foley, T., & Gurakar, M. (2022). Backlash or Bullying? Online Harassment, Social Sanction, and the Challenge of COVID-19 Misinformation. Journal of Online Trust and Safety, 1(2), Article 2. https://doi.org/10.54501/jots.v1i2.31	Communications	Public user messages posted on platform from users attending one of six large public universities (one in the southwest, two in the southeast, and three in the Midwest); also studied Platform's publicly accessible moderation policy	Observational and Theoretical/Conceptual	The authors' proposed framework, which evaluates user-generated content along dimensions of message intensity, how specific the message is to the target, and how persistent the messages are, improved accuracy of moderation activities when deployed on Patio.
Founta, A.-M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior (arXiv:1802.00393). arXiv. https://doi.org/10.48550/arXiv.1802.00393	Human Computer Interaction (HCI)	Recruited crowdsource annotators to label user-generated content referencing minority groups scraped from Twitter and then made statistical comparisons of the types of offensive speech annotated	Observational	This inquiry contributed an 80,000 Tweet manually annotated database identifying six distinct categories of online harassment material (offensive language, abusive language, hate speech, aggressive behavior, cyberbullying behavior, and spam).
Francisco, S. C., & Felmlee, D. H. (2022). What Did You Call Me? An Analysis of Online Harassment Towards Black and Latinx Women. Race and Social Problems, 14(1), 1–13. https://doi.org/10.1007/s12552-021-09330-7	Communications	Scraped public tweets containing deragatory language towards Black/Latinx posted 2015 - 2017 producing a final corpus of over 25,000 tweets	Observational	Online harassment reflects patterns of existing negative stereotypes surrounding race and gender. Knowledge of these patterns can inform platform policy.
Geiger, R. S. (2016). Bot-based collective blocklists in Twitter: The counterpublic moderation of harassment in a networked public space. Information, Communication & Society, 19(6), 787–803. https://doi.org/10.1080/1369118X.2016.1153700	Human Computer Interaction (HCI)	Discusses the history of block bots utilized on Twitter.	Theoretical/Conceptual	Block bots, in addition to being crowdsourced technical interventions, are also social interventions. The use of blockbots demonstrate the value of considering grassroots forms of moderations to support platforms for diverse user populations.
Georgakopoulos, S. V., Tasoulis, S. K., Vrahatis, A. G., & Plagianakos, V. P. (2018). Convolutional Neural Networks for Toxic Comment Classification. Proceedings of the 10th Hellenic Conference on Artificial Intelligence, 1–6. https://doi.org/10.1145/3200947.3208069	Human Computer Interaction (HCI)	Tested Convolution Neural Networks (CNN) approach vs. classic bag of words approaches on Wikipedia talk page comments	Observational and Theoretical/Conceptual	Cognitive Neural Networks (CNNs) improved text analysis for toxic comment classification. Future research should continue to examine applicability of this technique in social media contexts.

Gillespie, T. (2018). Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. Yale University Press.	Human Computer Interaction (HCI)	Draws upon a variety of peer-reviewed and non peer-reviewed empirical studies, policy documents, internal communications etc. to trace history of trust and safety work and the rise of modern social media platforms	Theoretical/Conceptual	This book explores the current infrastructure of content moderation including how policies are developed and implemented and impacts on social life. The major thesis of the book is that content moderation practices should receive more critical attention in the public sphere to better understand and recognize content moderation's important impacts on culture and well-being.
Golbeck, J., Ashktorab, Z., Banjo, R. O., Berlinger, A., Bhagwan, S., Buntain, C., Cheakalos, P., Geller, A. A., Gergory, Q., Gnanasekaran, R. K., Gunasekaran, R. R., Hoffman, K. M., Hottle, J., Jienjitlert, V., Khare, S., Lau, R., Martindale, M. J., Naik, S., Nixon, H. L., … Wu, D. M. (2017). A Large Labeled Corpus for Online Harassment Research. Proceedings of the 2017 ACM on Web Science	Human Computer Interaction (HCI)	Drawing upon an undesirable content glossary and the researchers' experiences, qualitatively coded a 35,000 tweet dataset (approximately 15% harassment examples and 85% negative interaction examples, per their evaluation criteria) for future researchers to use	Observational	This inquiry contributes a 35,000 tweet codebook which identifies linguistic and contextual features of online harassment for usre in future research projects.
Goyal, N., Park, L., & Vasserman, L. (2022). ”You have to prove the threat is real”: Understanding the needs of Female Journalists and Activists to Document and Report Online Harassment. CHI Conference on Human Factors in Computing Systems, 1–17. https://doi.org/10.1145/3491102.3517517	Human Computer Interaction (HCI)	Interviews and focus groups with 27 female journalists, activists, NGO employees	Interview and Design	Using PMCR (Prevention, Monitoring, Crisis and Recovery) framework along with the interview study, proposed a tool for target's to document harassment and reach out to support. Targets using the tool identified potential future directions to help targets be more likely to support violence and develop coping strategies.
Herring, S. C. (1999). The Rhetorical Dynamics of Gender Harassment On-Line. The Information Society, 15(3), 151–167. https://doi.org/10.1080/019722499128466	Communications	Observation of two interactions on an IRC (#india, 326 messages, channel claims mostly consiting of undergraduate students) and an academic listserv	Observational	Based on observations from these two online communities, women are excluded and silenced through language-based interpersonal harassment, threats, and violence. Women resisting this power differential are criticized invoking ideas of freedom of speech and censorship.
Huertas-García, Á., Martín, A., Huertas-Tato, J., & Camacho, D. (2023). Countering malicious content moderation evasion in online social networks: Simulation and detection of word camouflage. Applied Soft Computing, 145, 110552. https://doi.org/10.1016/j.asoc.2023.110552	Human Computer Interaction (HCI) and Communications	Tested analytical technique to overcome lexical content camoflaging techniques on a corpus of: OPUS news commentaries in 12 languages; 150,000 website domains across 23 languages captued with ParaCrawl; 4,000 TED talk transcripts from July 2020 translated into 100+ languages by volunteers; and parallel text	Observational and Theoretical/Conceptual	The pyleetspeak Python package is a starting point to use in content moderation to track how language evolves and is camoflaged to avoid detection mechanisms. The authors found that the model is most effective when offensive content is the noun or verb of the sentence and struggles with less central features such as pronouns and articles.
Im, J., Schoenebeck, S., Iriarte, M., Grill, G., Wilkinson, D., Batool, A., Alharbi, R., Funwie, A., Gankhuu, T., Gilbert, E., & Naseem, M. (2022). Women’s Perspectives on Harm and Justice after Online Harassment. Proceedings of the ACM on Human-Computer Interaction, 6(CSCW2), 355:1-355:23. https://doi.org/10.1145/3555775	Human Computer Interaction (HCI)	Survey in 14 geographic regions around the world (N = 3,993), with a fopcus on on regions whose perspectives have been historically neglected and overlooked in social media governance decisions (e.g. Mongolia, Cameroon)	Survey	Women experience online harm differently than men and want platforms to approach responding to online harms, especially non-consensual image sharing, differently as well. Women rate all forms of platform response more desirable than no response, and view revealing identities and bannning users most favorably.
Jhaver, S., Chen, Q. Z., Knauss, D., & Zhang, A. X. (2022). Designing Word Filter Tools for Creator-led Comment Moderation. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 1–21. https://doi.org/10.1145/3491102.3517505	Human Computer Interaction (HCI)	19 content creators on YouTube, Reddit, Twitch, TikTok and/or Instagram	Interview	In testing our FilterBuddy with social media content creators, the authors' identified several considerations for future interations including more transparency in outputs of word filters, ability to import word filters and other preferences, and improved organization.
Jhaver, S., Ghoshal, S., Bruckman, A., & Gilbert, E. (2018). Online Harassment and Content Moderation: The Case of Blocklists. ACM Transactions on Computer-Human Interaction, 25(2), 12:1-12:33. https://doi.org/10.1145/3185593	Human Computer Interaction (HCI)	14 Twitter users who used a popular blockbot, 14 users who were blocked by this blockbot	Interview	Blockbot users use blockbots to protect themselves online and engage with the platform in a safe, positive way, but all targets of blockbots believed they were blocked unfairly. The authors encourage, based on the contextual, personal nature of defining harassment, future research attention towards vulnerable groups and making blocklists more customizable and easier to interpret.

Jones, L., Mitchell, K., & Finkelhor, D. (2013). Online Harassment in Context: Trends From Three Youth Internet Safety Surveys (2000, 2005, 2010). Psychology of Violence, 3, 53–69. https://doi.org/10.1037/a0030309	Communications	Surveys of 4,561 youth Internet users (age 10 to 17) conducted in 2000, 2005, and 2010	Survey	Online harassment incident rates towards girl increased approximately 20% between 2000 and 2010. As more social dynamics move online, schools and other community stakeholders would benefit from increased efforts towards mitigation and response to cyberbullying.
Kou, Y. (2020). Toxic Behaviors in Team-Based Competitive Gaming: The Case of League of Legends. Proceedings of the Annual Symposium on Computer-Human Interaction in Play, 81–92. https://doi.org/10.1145/3410404.3414243	Human Computer Interaction (HCI)	476 posts and 40,133 accompanying comments from League of Legends Reddit forum	Observational	The authors identified five types of toxic behavior: communicative aggression, cheating, hostage holding, mediocritizing, and sabotaging. Addressing all of these toxic behaviors require considering contextual factors motivating their use, including competitiveness and social status.
Kou, Y. (2021). Punishment and Its Discontents: An Analysis of Permanent Ban in an Online Game Community. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 334:1-334:21. https://doi.org/10.1145/3476075	Human Computer Interaction (HCI)	197 threads, consisting of 7,142, from a League of Legends Reddit forum	Observational	Platform bans in this community do not support users becoming more positive community members, but rather situates being a banned user as a rhetorical device to assert platform power and authority.
Kou, Y., Gui, X., Zhang, S., & Nardi, B. (2017). Managing Disruptive Behavior through Non-Hierarchical Governance: Crowdsourcing in League of Legends and Weibo. Proceedings of the ACM on Human-Computer Interaction, 1(CSCW), 62:1-62:17. https://doi.org/10.1145/3134697	Human Computer Interaction (HCI)	Observed the Tribunal in League of Legends' North America server from September 2011 to April 2014, interviewed a Riot Games designers and conducted 52 interviews with players and Tribunal judges	Observational and Interview	Though most community members appreciated the hierarchical governance approach implemented, many raised concerns with the effectiveness of the system, distrust about lack of transparency about the Tribunal process, and judges' desires to engage with the process beyond the capabilities of the current system.
Li, Q. (2005). Gender and CMC: A review on conflict and harassment. Australasian Journal of Educational Technology, 21(3), Article 3. https://doi.org/10.14742/ajet.1327	Communications	Literature review of education literature centered on gender and online conflict/harassment published between 1966 and 2002	Theoretical/Conceptual	Implementing technology in educational contexts must consider gender differences in communication and outcomes when using technology. The author argues that individuals may engaged with gendered identities and behaviors more fluidly online, requiring holisitic assessment of gender differences rather than assumed dynamics (i.e. female victims, male perpetrators).
Lwin, M. O., Li, B., & Ang, R. P. (2012). Stop bugging me: An examination of adolescents' protection behavior against online harassment. Journal of Adolescence, 35, 31–41.	Communications	Surveyed 537 adolescents/young adults aged 12 to 19 residing in Singapore	Survey	Perceived susceptibility does not predict individual youths' intentions to engage in protective strategies, suggesting that more education and awareness surrounding the severity of online harassment is important for youth populations in particular.
Mahar, K., Zhang, A. X., & Karger, D. (2018). Squadbox: A Tool to Combat Email Harassment Using Friendsourced Moderation. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1–13. https://doi.org/10.1145/3173574.3174160	Human Computer Interaction (HCI)	Interviewed 18 victims of online harassment, developed Squadbox application, tested Squadbox with 5 friends duos	Interview and Design	SquadBox was successful in that users trusted friends to engage in anti harassment efforts as opposed to strangers, but this approach raised tensions related to user privacy, moderaiton burdens on friends, and timeline pressures.
Malcorps, S., Libert, M., & Le Cam, F. (2022). A Matter of Organisational Silence: Media Managers Struggling to Make Sense of the Online Harassment of Journalists as a Collective Issue in Journalism. Digital Journalism, 0(0), 1–18. https://doi.org/10.1080/21670811.2022.2140301	Communications	Interviews and focus groups with 22 Belgian media managers	Interview	There is not a consistent definition of online harassment or policy/protocol for response across media organizations.

Matias, J. N. (2019). Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. Proceedings of the National Academy of Sciences, 116(20), 9785-9789. https://doi.org/10.1073/pnas.1813486116	Human Computer Interaction (HCI)	Examined effect of new rule postings on Reddit r/science forum from August 25, 2016 through September 23, 2016. Examined 2,190 discussions and identified 18,264 newcomer comments in 804 discussions.	Observational and Experimental	Implementing random reminders of community guidelines in popular discussions predicted higher rates of positive behavior, especially for new members to the community.
McFarlane, L., & Bocij, P. (2003). An exploration of predatory behaviour in cyberspace: Towards a typology of cyberstalkers. First Monday. https://doi.org/10.5210/fm.v8i9.1076	Communications	24 interviews with cyberstalking victim survivors	Interview	The authors identified four major archetypes of cyberstalking: vindictive, composed, intimate, and collective. These four archetypes have varying degrees of overlap with traditional stalking, and future research is needed to examine prevention and mitigation effort effectiveness.
McGraw, D. K. (1995). Sexual Harassment in Cyberspace: The Problem of Unwelcome E-mail Note. Rutgers Computer & Technology Law Journal, 21(2), 491–518.	Communications	Examines the legal, cultural, and technical context of sexual harassment via email and implications for women	Theoretical/Conceptual	Current legal infrastructure and cultural dynamics disadvantage womens' safe and equitable access to email communications. Future policy, research, and design needs to critically consider the how email contributes to gender inequalities.
Moneva, A., Miró-Llinares, F., & Hart, T. C. (2021). Hunter or Prey? Exploring the Situational Profiles that Define Repeated Online Harassment Victims and Offenders. Deviant Behavior, 42(11), 1366–1381. https://doi.org/10.1080/01639625.2020.1746135	Communications	4174 surveys with Spanish speaking non-university students aged 12 to 21; analyzed 10 variables, including individual-level characteristics and self-reported online behavior, to develop predictive profiles of online harassment victimization and offending	Survey	The forms of online harassment identified in this project (archetypes identified based upon frequency of harassment, demographics, time spent online etc.) occur in distinct social and situational contexts, requiring platform policy to consider and respond differently based upon the scenarios associated with different forms of harassment.
Moore, R., Guntupalli, N. T., & Lee, T. (2010). Parental Regulation and Online Activities: Examining factors that influence a Youth's potential to become a Victim of Online Harassment. International Journal of Cyber Criminology, 4(1/2), 685–698.	Communications	935 surveys collected from October to November 2006 as part of the Pew Internet and American Life project	Survey	Parental regulation did not have a strong predictive relationship with youths' likelihood of experiencing online harassment, suggesting more research is needed to fully understand the role of parentla engagement in online harassment.
Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. New Media & Society, 20(11), 4366–4383. https://doi.org/10.1177/1461444818773059	Human Computer Interaction (HCI)	389 user reports of personal experiences of content takedown on Twitter, Instagram, YouTube, and Google+	Survey	User folks models of online content moderation are often incomplete or inaccurate, suggesting education as a valuable and underutilized strategy for preventing undesirable behavior online.
Nadim, M., & Fladmore, A. (2019). Silencing Women? Gender and Online Harassment. Social Science Computer Review, 39(2), 245–258.	Human Computer Interaction (HCI)	Two surveys with residents of Norway (2013 survey on experiences with unpleasant comments N = 1,534; 2016 survey on experiences with hateful comments N = 5,054)	Survey	Gender differences in prevalence rates and experiences of online harassment are related to the topic harassing comments are targeted at and if harassment content is related to group characteristics or not. Understanding and preventing online harassment requires capturing the context and content of messages in relation to potential gender differences.
Näsi, M., Räsänen, P., Oksanen, A., Hawdon, J., Keipi, T., & Holkeri, E. (2014). Association between online harassment and exposure to harmful online content: A cross-national comparison between the United States and Finland. Computers in Human Behavior, 41, 137–145. https://doi.org/10.1016/j.chb.2014.09.019	Human Computer Interaction (HCI)	2013 survey with individuals aged 15 to 30 in US (N=1,032) and Finland (N=555)	Survey	The authors found associations between experiencing online harassment and spending time on websites related to eating disorders, age, gender, and subjective well-being scores. Similarities across US and Finland samples suggest similar approaches to online harassment ay be effective across cultural contexts.

Ozanne, M., Bhandari, A., Bazarova, N. N., & DiFranzo, D. (2022). Shall AI moderators be made visible? Perception of accountability and trust in moderation systems on social media platforms. Big Data & Society, 9(2), 20539517221115666. https://doi.org/10.1177/20539517221115666	Human Computer Interaction (HCI)	582 MTurkers participated in a social media simulation content moderation experiment	Experiment	Users perceived AI-based moderation decisions as less accountable and less trustworthy than human made decisions, but there were no statistical differences in perceived moderation fairness or participant knowledge of the moderation process. These findings suggest implementing AI backed content moderation mechanisms will require unique considerations for user perceptions of
Page, X., Knijnenburg, B. P., Wisniewski, P., & Namara, M. (2018). Avoiding Online Harassment: The Socially Disenfranchised. In J. Golbeck (Ed.), Online Harassment (pp. 243–268). Springer International Publishing. https://doi.org/10.1007/978-3-319-78583-7_11	Human Computer Interaction (HCI)	Literature review of technology non-use resulting from experiences of online harassment along with 17 interviews with adult self-identified social media non-users	Interview and Theoretical/Conceptual	The potential negative social consequences of technology non-use can be mitigated by including features for non-users to access, content, provide ways for non-users to flag or have content removed, and include opportunities for sharing and connection through other modalities.
Pater, J. A., Kim, M. K., Mynatt, E. D., & Fiesler, C. (2016). Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms. Proceedings of the 19th International Conference on Supporting Group Work, 369–374. https://doi.org/10.1145/2957276.2957297	Human Computer Interaction (HCI)	Reviewed publicly accessible policy and platform governance documentation (a total of 56 documents) from 15 social media platforms	Theoretical/Conceptual	Based upon this document review, there is inconsistency in how platforms define harassment and how enforcement techniques are determined and implemented.
Phillips, A. L. (2018). Youth Perceptions of Online Harassment, Cyberbullying, and “just Drama”: Implications for Empathetic Design. In J. Golbeck (Ed.), Online Harassment (pp. 229–241). Springer International Publishing. https://doi.org/10.1007/978-3-319-78583-7_10	Human Computer Interaction (HCI)	8 semi-structured interviews	Interview	Youth hesitate to describe online harassment as bullying or report it to formal support services due to stigma associated with bullying and victimization. The authors recommend more cyberbullying resources and education for responding after cyberbullying has begun and design systems which center empathy with the experiences of peers.
Pilipets, E., & Paasonen, S. (2022). Nipples, memes, and algorithmic failure: NSFW critique of Tumblr censorship. New Media & Society, 24(6), 1459–1480. https://doi.org/10.1177/1461444820979280	Human Computer Interaction (HCI)	7,306 Tumblr posts with #censorship tag published between November 2018 and April 2019	Observational	In response to Tumblr's new policy and algorithmic flagging protocol for nudity, community members expressed dissatisfaction and criticism towards the brand via cultural production including memes and other collective sharing as a form of protest. These findings re-iterate the importance of a shared understanding between platforms and users towards undesirable content (in this case, NSFW content) for effective policy change and community buy-in.
Raisi, E., & Huang, B. (2017). Cyberbullying Detection with Weakly Supervised Machine Learning. Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, 409–416. https://doi.org/10.1145/3110025.3110049	Human Computer Interaction (HCI)	Built machine learning model to identify bullies on Twitter and predict future bullying content from individual users	Design	This model demonstrated improved accuracy in predicting cyberbullies and victims based on context in three social media datasets, suggesting potential efficacy in applied content moderation work.
Razi, A., Kim, S., Alsoubai, A., Stringhini, G., Solorio, T., De Choudhury, M., & Wisniewski, P. J. (2021). A Human-Centered Systematic Literature Review of the Computational Approaches for Online Sexual Risk Detection. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 465:1-465:38. https://doi.org/10.1145/3479609	Human Computer Interaction (HCI)	Literature review of 73 empirical peer-reviewed articles using computational approaches to detect online sexual risk	Theoretical/Conceptual	Most work on detection of sexual content targeting youth identified content after it was produced (and thus, reached victims), was generated from public datasets, and focused on algorithmic performance rather than evaluation with actual users. Future research approaches would benefit from expanded focus on prevention and insights from real world users.
Sanders, T., Trueman, G., Worthington, K., & Keighley, R. (2023). Non-consensual sharing of images: Commercial content creators, sexual content creation platforms and the lack of protection. New Media & Society, 14614448231172711. https://doi.org/10.1177/14614448231172711	Human Computer Interaction (HCI)	221 survey responses and 16 interviews with content creators	Survey and Interview	Online sex workers face barriers from user harassment and platform policies. In responding to image-based harassment these content creators face, these creators recommend more inclusion of their perspectives in the bottom-up development of platforms and strong legal pathways for abuse of sexual imagery.

Schenk, S. (2008). Cyber-Sexual Harassment: The Development of the Cyber-Sexual Experiences Questionnaire. Journal of Computer-Mediated Communication, 12.	Communications	Focus groups of 2 to 6 individuals with 24 undergraduate women	Interview	The 21-item Cyber-Sexual Experiences Questionnaire can be used to capture experiences of sexual harassment in online spaces and contribute to future research across identity groups and online contexts.
Scheuerman, M. K., Jiang, J. A., Fiesler, C., & Brubaker, J. R. (2021). A Framework of Severity for Harmful Content Online. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 368:1-368:33. https://doi.org/10.1145/3479512	Human Computer Interaction (HCI)	Interviews and card-sorting exercises with 52 participants	Interview	While only about half of participants explicitly reported experiencing online sexual harassment, per the study definition, every participant experienced this form of harassment, suggesting further education on forms online harassment would be valuable. Additionally, most harassment incidents took place on social media and instant messenger systems, suggesting future research on particular risks of these forms of communication.
Schoenebeck, S., Batool, A., Do, G., Darling, S., Grill, G., Wilkinson, D., Khan, M., Toyama, K., & Ashwell, L. (2023). Online Harassment in Majority Contexts: Examining Harms and Remedies across Countries. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 1–16. https://doi.org/10.1145/3544548.3581020	Human Computer Interaction (HCI)	4,000 survey respondents from 14 different countries	Survey	The four major forms of harm and eight dimensions of harm identified by this study can be applied to research and practice to better understand user experiences of harm and implement effective policy.
Schulenberg, K., Li, L., Freeman, G., Zamanifard, S., & McNeese, N. J. (2023). Towards Leveraging AI-based Moderation to Address Emergent Harassment in Social Virtual Reality. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 1–17. https://doi.org/10.1145/3544548.3581090	Human Computer Interaction (HCI)	Interviews with 39 social virtual reality users	Interview	Users identified technical limitations to transplanting existing methods of content moderation to VR settings to novel forms of harassment. The authors identify collaborations with users, spreading moderation decisions across users and AI, and recruiting diverse human moderators as approaches to address these challenges.
Seering, J., Dym, B., Kaufman, G., & Bernstein, M. (2022). Pride and Professionalization in Volunteer Moderation: Lessons for Effective Platform-User Collaboration. Journal of Online Trust and Safety, 1(2), Article 2. https://doi.org/10.54501/jots.v1i2.34	Human Computer Interaction (HCI)	Semi-structured interviews with 56 volunteer moderators on Twitch, Reddit, and Facebook	Interview	Platform collaborations with volunteer moderators can be an effective approach, but impelementation should consider ethical considerations, consider the diverse forms volunteer collaborations could take, and allocate sufficient resources to support the success of volunteer moderators.
Seering, J., Wang, T., Yoon, J., & Kaufman, G. (2019). Moderator engagement and community development in the age of algorithms. New Media & Society, 21(7), 1417–1443. https://doi.org/10.1177/1461444818821316	Human Computer Interaction (HCI)	11 moderators working in LGBTQ+ Discord communities	Interview	Volunteer moderators faced identity-based harassment, including hate speech and threats, in their roles, raising ethical considerations and the importance of platforms considering how to support and protect volunteer mods. Additionally, the authors argue that this approach was successful because of Discord building trust and formal networks of contact and support with the recruited volunteers.
Sheridan, L. P., & Grant, T. (2007). Is cyberstalking different? Psychology, Crime & Law, 13(6), 627–640. https://doi.org/10.1080/10683160701340528	Communications	1,051 stalking victims residing in the US, UK, and Australia	Survey	Cyberstalking reflects many elements of traditional stalking, suggesting that applying effective techniques from traditional stalking contacts could be effective in prevention and response.

Suzor, N., Dragiewicz, M., Harris, B., Gillett, R., Burgess, J., & Van Geelen, T. (2019). Human Rights by Design: The Responsibilities of Social Media Platforms to Address Gender-Based Violence Online. Policy & Internet, 11(1), 84–103. https://doi.org/10.1002/poi3.185	Human Computer Interaction (HCI)	Conceptual discussion of the current state of gender-based violence and recommendations for social media platforms and policy actors	Theoretical/Conceptual	Governments are recommended to utilize existing international human rights instruemnts in outlining responsibilities of platforms in addressing gender-based violence. Recommendations from the authors include more empirical research to understand the state of online GBV, especially centered on victim impacts, and continued collaborations between governments and civil society stakeholders to impact policy decisions.
Tolba, M., Ouadfel, S., & Meshoul, S. (2021). Hybrid ensemble approaches to online harassment detection in highly imbalanced data. Expert Systems with Applications, 175, 114751. https://doi.org/10.1016/j.eswa.2021.114751	Human Computer Interaction (HCI)	Development of a harassment detection model using a Twitter dataset	Design	This study's approach of multiple computational techniques, intended to more accurately capture harassment in real social media datasets, has evidenced effectiveness as compared to other techniques, demonstrating potential value for future computational research in harassment detection.
Uttarapong, J., Cai, J., & Wohn, D. Y. (2021). Harassment Experiences of Women and LGBTQ Live Streamers and How They Handled Negativity. ACM International Conference on Interactive Media Experiences, 7–19. https://doi.org/10.1145/3452918.3458794	Human Computer Interaction (HCI)	Semi-structured interviews with 25 women and/or LGBTQ+ Twitch streamers	Interview	Streamers, in addition to technical challenges in moderating their channels, bear the brunt of emotional and social labor in managing their communities. The authors recommend the support of third spaces for women and LGBTQ+ streamers to connect about their experiences, verification mechanisms to prevent harassers in real time channels, and the creation of policy guidelines mental health support and other resources.
Vitak, J., Chadha, K., Steiner, L., & Ashktorab, Z. (2017). Identifying Women's Experiences With and Strategies for Mitigating Negative Effects of Online Harassment. Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, 1231–1245. https://doi.org/10.1145/2998181.2998337	Human Computer Interaction (HCI)	659 surveyered undergraduate and graduate student women	Survey	To support women users' harassment mitgation strategies, the authors suggest creating more user accessible custom filtering tools, offering alternatives to change engagement/information sharing other than deactivation, education for users about impression management and data privacy strategies, and quicker platform responses to harassment.
Wei, M., Consolvo, S., Kelley, P. G., Kohno, T., Roesner, F., & Thomas, K. (2023). "There's so much responsibility on users right now:" Expert Advice for Staying Safer From Hate and Harassment. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 1–17. https://doi.org/10.1145/3544548.3581229	Human Computer Interaction (HCI)	Interviews with 24 online safety experts (12 academics, 7 NGO employees and 6 industry professionals) about combatting online harm	Interview	There is considerable disagreement amongst best practices for avoiding online harassment from professionals, but consensus did emerge surrounding recommendations including: muting and blocking are valuable but raise freedom of expression challenges, there should be restrictions on publicly accessible material, impersonation is an issue with no good solutions, and there is a lack of effective strategies for users experiencing overloading.
Wohn, D. Y. (2019). Volunteer Moderators in Twitch Micro Communities: How They Get Involved, the Roles They Play, and the Emotional Labor They Experience. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–13. https://doi.org/10.1145/3290605.3300390	Human Computer Interaction (HCI)	Interviews with 20 Twitch moderators	Interview	Users were driven to become moderators via the necessity of their streamers and there was stigma around expressing a desire to be a mod. Moderators primarily curating content and engaging in complex social dynamic management on the channel and between the mod and streamer, and future research on moderation should more critically examine these dynamics.
Xiao, S., Jhaver, S., & Salehi, N. (2023). Addressing Interpersonal Harm in Online Gaming Communities: The Opportunities and Challenges for a Restorative Justice Approach. ACM Transactions on Computer-Human Interaction. https://doi.org/10.1145/3603625	Human Computer Interaction (HCI)	Interviews with 23 moderators, victims, and offenders in the Overwatch community along with observing communications on an Overwatch Discord channel	Interview and Observational	Content moderation systems in the Overwatch community face technical, social, and policy barriers to implementing restorative justice. The authors recommend applying restorative justice explanations to content moderation decisions, have conferences with victims/offenders, and deploy restorative justice techniques only when participants are willing to adhere to the process and it aligns with community values.

Yang, E., Lewis, D. D., & Frieder, O. (2021). TAR on Social Media: A Framework for Online Content Moderation (arXiv:2108.12752). arXiv. http://arxiv.org/abs/2108.12752	Human Computer Interaction (HCI)	Testing technology-assisted review (TAR) framework on social media datasets	Design	Technology-assisted review (TAR) can reduce human costs of moderation through routing duplicates and batched conversations to the same human reviewer, but testing in the complexities of real world social media scenarios is needed.
Ybarra, M. L., Diener-West, M., & Leaf, P. J. (2007). Examining the Overlap in Internet Harassment and School Bullying: Implications for School Intervention. Journal of Adolescent Health, 41(6), S42–S50. https://doi.org/10.1016/j.jadohealth.2007.09.004	Communications	Surveyed 1,588 10 to 15 year old English speakers residing in the US	Survey	Youth harassed online are more likely to face trouble in school, suggesting that schools could be an important site of support and education surrounding cyberbullying.
Ybarra, M. L., & Mitchell, K. J. (2004a). Online aggressor/targets, aggressors, and targets: A comparison of associated youth characteristics. Journal of Child Psychology and Psychiatry, 45(7), 1308–1316. https://doi.org/10.1111/j.1469-7610.2004.00328.x	Communications	Phone surveys with 1,501 Internet users 10 to 17 years old in the US	Survey	Youth experience high rates of online harassment and subsequent negative outcomes such as depresive symptoms. Interventions should consider considering youth perspectives as valuable in developing safety education and recognize that bullying happens offline and online. More research is needed to understand demographic trends in who experiences harassment and outcomes.
Ybarra, M. L., & Mitchell, K. J. (2004b). Youth engaging in online harassment: Associations with caregiver–child relationships, Internet use, and personal characteristics. Journal of Adolescence, 27(3), 319–336. https://doi.org/10.1016/j.adolescence.2004.03.007	Communications	Phone surveys with 1,501 parent child dyads in the US	Survey	Poor parent-child relationships predict being a cyberbully online. Interventions would benefit from considering overlap with traditional bullying risk factors.
Ybarra, M. L., & Mitchell, K. J. (2008). How Risky Are Social Networking Sites? A Comparison of Places Online Where Youth Sexual Solicitation and Harassment Occurs. Pediatrics, 121(2), e350–e357. https://doi.org/10.1542/peds.2007-0693	Communications	Phone surveys with 1,501 Internet users 10 to 17 years old in the US	Survey	Youth online sexual harassment may not be as widespread an issue as current risk messaging suggests. Prevention efforts may be more effective if they focus on social problems youth face broadly such as mental health services.
Ybarra, M. L., Mitchell, K. J., Wolak, J., & Finkelhor, D. (2006). Examining Characteristics and Associated Distress Related to Internet Harassment: Findings From the Second Youth Internet Safety Survey. Pediatrics, 118(4), e1169–e1177. https://doi.org/10.1542/peds.2006-0815	Communications	Surveyed 1588 youth age 10 to 15 years old in the US	Survey	Youth bullyed in other contexts, who harassed others, who faced social challenges, and who used personal messaging functionalities were more likely to be victims of cyberbullying. Youth experiencing harassment from an adult were three times more likey to report distress. Intervention programs should center on developing youth interpersonal skills and antibullying education generally, in addition to developing standards for platforms to be proactive in addressing serious incidents of harms alongside users.