Commentary

Bridging Theory & Practice: Examining the State of Scholarship Using the History of Trust and Safety Archive

Megan Knittel and Amanda Menking

1 Introduction

In June 2023, Casey Newton, an American technology journalist, asked "Have we reached peak trust and safety?" (2023).[1] Newton was responding rhetorically to a particular moment in time, one in which platforms seemed to be reducing investment in trust and safety overall, laying off entire teams (for example, after Elon Musk took ownership of the social media platform formerly known as Twitter), relaxing policies (Newton 2023), losing the trust of unpaid community moderators (Nix 2023; Roose et al. 2023), and slashing budgets for internal research (Vanian and Field 2023). However, recent world events—including but not limited to the COVID-19 pandemic, rising rates of tech abuse and online harassment, and mis/disinformation campaigns from politically motivated actors—have prompted increased attention in peer-reviewed scholarship about trust and safety topics.

In particular, growing attention in scholarly spaces has centered on understanding online hurdles for human rights in our current digitally mediated information landscape (Suzor et al. 2019), particularly for historically marginalized communities (Noble and Tynes 2016; Zolides 2021). The examples above demonstrate a growing recognition across the social sciences that "trust and safety" is an important research topic—even if scholars do not use this term.

Beyond a general recognition of broad trends in how social scientists engage with the challenges of creating equitable digital landscapes, however, no attempt has been made to quantify or otherwise describe how peer-reviewed scholarship has discussed trust and safety. It is unclear how topics of interest (and the associated terminology) relevant to trust and safety (T&S) professionals are represented in peer-reviewed academic scholarship. To this end, if and how academic investigations intersect with T&S practice remains unclear. Our project asks: *How does peer-reviewed scholarship approach issues relevant to the practice of trust and safety?*

The precedent of disconnect between peer-reviewed scholarship and on-the-ground T&S practice disadvantages both T&S professionals and researchers. This commentary attempts to bridge the lack of current engagement between T&S professionals and academic scholars by introducing the History of Trust and Safety Archive, a collaborative, dynamic Zotero database of peer-reviewed scholarship relevant to trust and safety (TSF

1. Newton's work builds on Kate Klonick's analysis of the "Golden Age of Tech Accountability" (Klonick 2023). In response to a panelist's prompt, "Was 2021, in retrospect, a heyday for trust and safety online?," Klonick unpacks the last five years of big tech regulation and industry trends.

10.54501/jots.v2i2.173

2023). Our goal is for the History of Trust and Safety Archive to serve as a platform for connection and collaboration both academics and T&S professionals can use to learn from one another and develop a shared foundation of knowledge.

This commentary summarizes our findings from a semi-systematic literature review using the Archive, and is a pilot test of using the Archive for a focused research purpose. We focus on a narrow slice of the peer-reviewed trust and safety scholarship in the History of Trust and Safety Archive: Communications and Human Computer Interaction (HCI) scholarship. We used the keywords "harassment" and "cyberbullying" to query the Archive. We selected these keywords to reflect both the glossaries of trust and safety terms available to the public and the language used in peer-reviewed scholarship. We chose to focus on the specific abuse types "harassment" and "cyberbullying" because these terms are recognized by T&S professionals and are also frequently used in Communications and HCI scholarship. Additionally, we selected these terms and associated abuse types to capture different dimensions of undesirable and illegal content.

In our review, we found that peer-reviewed trust and safety scholarship follows several trends. First, the majority of work focused on harms and problems with online content and, subsequently, on developing techniques for identifying and preventing undesirable content. Additionally, the datasets researchers used primarily focused on publicly accessible social media data and the experiences of volunteer moderators and community members. Based on these patterns, the core takeaways of this commentary are as follows.

First, a takeaway for researchers is the need for increased focus on collaborations with T&S professionals to capture their valuable perspectives in peer-reviewed scholarship. Future research would benefit from expanding the topics and populations studied.

Second, the major takeaway from this project for T&S professionals is that the dynamic challenges of their day-to-day T&S work, such as responding to new problems and the complications of real-world datasets, are not well represented in the peer-reviewed canon. Facilitating points-of-knowledge sharing and collaboration between scholars and professionals will be critical for producing research that accurately captures T&S as a phenomenon, practice, and profession.

2 Approach

Our approach to this literature review was both semi-systematic and iterative. To evaluate the current state of peer-reviewed scholarship, we developed an archive of peer-reviewed scholarship relevant to trust and safety.[2] Focusing our inquiry on peer-reviewed manuscripts/articles and monographs/books published by university presses, this search and manual review resulted in 1,287 items tagged as relevant to five social science disciplines: Communications, Human Computer Interaction (HCI), Law and Policy, Media Studies, and Science and Technology Studies (STS). Complementing the keyword

2. We began to develop the Archive by formulating the problem: a lack of work reviewing and describing peer-reviewed contributions on T&S topics. From there, we identified broad project goals and objectives. These goals were developed through reflection and discussion with the Trust and Safety Foundation (TSF) team including both co-authors and community members, most notably Jeff Lazarus, a T&S professional who initially brought this project idea to TSF. Lazarus identified a lack of research surrounding the history of trust and safety as a profession. The team agreed this deficit also included a lack of scholarship capturing the firsthand experiences of T&S professionals who had established the field across the tech sector. Consequently, TSF began the History of Trust & Safety Project, including an interview-based study with practitioners and others in the T&S ecosystem. This manuscript is a more tightly scoped investigation within the broader project. In this manuscript, we are specifically focused on exploring the peer-reviewed academic literature.

approach, we recruited subject matter experts (SMEs) to review the Archive and provide their feedback. All of these experts, drawn from diverse contexts including academia, nongovernmental organizations, industry, and policy spaces, were intentionally selected based on their experience with trust and safety.

For this literature review using the Archive, we focus on HCI and Communications scholarship, because these disciplines represent a high degree of relevant activity in the Archive. We used the keywords "harassment" and "cyberbullying" to query the Archive. These keywords were selected based on their widespread use across formal and informal trust and safety contexts; they also describe abuse types in terms T&S professionals use (TSPA, n.d.).

3 Research Questions

Virtually no attempts have been made to describe how trust and safety issues, like platform governance and online harassment, have been studied across academic disciplines. As noted above, as a starting point, we started by comparing HCI and Communications scholarship with regard to "harassment" and "cyberbullying," leading to Question 1:

1. How does HCI and Communications peer-reviewed academic scholarship define online harassment and cyberbullying?

Next, an intended application of the Archive is for others to be able to identify which methodological approaches have been used to understand trust and safety topics. Understanding the current state of scholarship includes critically evaluating the quality of the data that produces the findings used to inform future research. Our second research question centers on summarizing how HCI and Communications study "harassment" and "cyberbullying":

2. How does HCI and Communications peer-reviewed academic scholarship operationalize online harassment and cyberbullying?

Peer-reviewed scholarship develops and changes over time. Particularly in the context of new technologies and the social changes accompanying them, scholarship varies in its level of attention to different topics. Capturing these trends is critical for identifying potential oversights and opportunities for new work, prompting Research Question 3:

3. What trends in understanding online harassment and cyberbullying are represented in HCI and Communications peer-reviewed academic scholarship?

Finally, a major goal of this manuscript and the Archive is to connect the valuable insights of peer-reviewed scholarship to the work of T&S professionals. We developed our fourth research question to identify areas of connection, and potential opportunities for expansion, between scholarship and the experiences of professionals:

4. How does peer-reviewed HCI and Communications academic scholarship present itself as informing trust and safety work from the perspective of professionals in the field?

4 Findings

Across the 75 peer-reviewed articles and books comprising this review, 45 (60%) collected data directly from users or studied user-generated content (UGC). This included any kind of user population from country-level surveys with school-aged youth (Jones,

Mitchell, and Finkelhor 2013; Ybarra et al. 2006) to interviews with individuals who have experienced severe online harassment (Blackwell et al. 2017; Chadha et al. 2020). Ten manuscripts (approximately 13% of items in the review) focused on populations of users who directly contributed to content moderation on social media platforms in some way. These community moderator-focused contributions included examinations of volunteer moderators in online communities (Seering and Kairam 2022; Wohn 2019) and several experiments that placed platform users into the role of content moderators (Foley and Gurakar 2022). The remaining 20 studies (approximately 27%) used empirical methods that did not study users or UGC. Of these 20 non-user studies, three were primarily focused on reviewing and describing publicly posted platform policies related to harassing users, inflammatory content, and other trust and safety concerns.

From this overview, we noted what the research did *not* discuss. For example, none of the studies gathered empirical data consisting of social media content (i.e., forum comments, social media posts, images, private messages, etc.) that had been removed or modified through content moderation. Several studies used language keywords to identify harassing language in scraped UGC social media datasets (Dinakar et al. 2012; Felmlee et al. 2020; Founta et al. 2018; Francisco and Felmlee 2022; Golbeck et al. 2017; Raisi and Huang 2017). These studies used lists of undesirable language (including derogatory terms related to racial and ethnic groups, sexual minorities, and women) to evaluate the frequency of this content across time and context. However, none of the studies in our literature review examined content that was publicly accessible in some way. Researchers did not include any UGC that had been removed or altered, such as through automated flagging and content removal policies, prior to being captured.

5 Focus on Harms and Problems

The first major finding from this systematic review is that HCI and Communications scholarship relevant to trust and safety primarily centers on harms and problems online. Virtually every manuscript identified the challenges online platforms face in protecting users and preventing harmful and undesirable UGC, and the harms that targets of online harassment face.

The manuscripts in this literature review capture a wide range of topics, contexts, study populations, platforms of interest, and theoretical orientations to understand online harassment. As a result, there is not any one particular definition or criteria used to determine what is or is not online harassment. Studies operationalized harassment and bullying primarily via UGC coding based on keyword lists of identity-based harassment and bullying content (Golbeck et al. 2017) and/or user-identified interpersonal harassment (Goyal, Park, and Vasserman 2022). Even in more topically focused investigations, such as Feng and Wohn's examination of bystander interventions in online harassment, we found that the terms "harassment" and "cyberbullying" were used interchangeably across studies (Feng and Wohn 2020).

Across these differences, two overarching themes about what content does or does not constitute online harassment emerged. First, UGC that constitutes online harassment is generally content that relates to some kind of adverse outcome(s), such as adverse mental health outcomes or users not wanting to use the online platform anymore (Page et al. 2018). The second unifying thread across approaches to online harassment was that the harassing content had a clear target, either an individual (Blackwell et al. 2017; Golbeck et al. 2017), a social identity group(s) (Francisco and Felmlee 2022), or potentially both (Duguay, Burgess, and Suzor 2020). Online harassment was distinguished from other types of harm in UGC via connecting harassing content to humans, individuals,

or groups, in a negative or aggressive way. Both these trends are incredibly broad but serve as a starting point for connecting online harassment concepts across field sites, populations, and methodologies.

6 Identifying and Preventing Undesirable Content

The second major finding was that the scholarship we reviewed focused on developing technical approaches and theoretical frameworks to identify and prevent undesirable content online. A trend across the Communications and HCI scholarship was seeking an understanding of how to address undesirable content more effectively and accurately, and specifically how to prevent undesirable UGC.

Many research contributions focused on describing how much online harassment content was produced on specific platforms during certain periods of time, such as in the wake of platform policy decisions (DeCook et al. 2022), or in databases of UGC scraped and qualitatively coded to identify harassing content (Golbeck et al. 2017). A handful of publications, rather than collecting new empirical data, reviewed other academic scholarship to identify trends in prevalence rates and forms of abuse (Razi et al. 2021).

7 The Perspectives of Users vs. Platforms

The final major finding from this review is that the experiences of users are portrayed in opposition to platforms. The experiences of users, particularly negative experiences, are highlighted to demonstrate perceived oversights and problems with listed policy and platform governance documentation. A majority of investigations in this review centered on observing user data, mostly text-based posts on social media platforms (Chatzakou et al. 2017) or interviews (Jhaver et al. 2022) and surveys (Im et al. 2022) with users.

Platform users often have an outsiders' view of how content is subject to moderation policy. An emerging thread of research focused on how users perceive content that has been subject to perceived moderation activity (i.e., user responses to flagged content [Bhandari et al. 2021] or what users say about the efficacy and transparency of moderation activities [Myers West 2018; Sanders et al. 2023]). These investigations identified a pattern of users and content creators feeling a disconnect between posted policy and platform rules and consequences. For example, one study examined the use of third-party blocklist applications that work with the Twitter API as an example of "bottom-up," user-driven efforts to control the material they see and the people they interact with on Twitter (Jhaver et al. 2018).

8 The Future of Trust & Safety Research

This literature review examined how peer-reviewed scholarship from Communications and HCI discussed "harassment" and "cyberbullying." In the 75 documents we reviewed, we identified three major findings relevant to T&S professionals and academic scholars working on trust and safety topics. First, academic inquiries centered on characterizing experiences of harm from online platforms. Second, the research focused on identifying harmful content and developing technical and procedural strategies for preventing and responding to it. Finally, this body of research prioritized the perspectives of users and often portrayed them as in conflict with that of platforms.

In addition to the trends we identified, this investigation also contributes an intellectual

and practical framework for conducting literature reviews, using whatever trust and safety topics and keywords desired, to ask and answer questions with the Archive. From this review, there are several takeaways for T&S practitioners to consider when supporting their teams, connecting with academic scholars, and developing ways to manage emerging trust and safety challenges.

9 Takeaways for T&S Practitioners

The first takeaway for T&S practitioners is that online harassment/cyberbullying scholarship in Communications and HCI almost exclusively considers public-facing data to understand trust and safety issues. Virtually all data is from an "outsider's" perspective via publicly accessible documentation, observable UGC, and the perceptions of everyday platform users. The investigations in the Archive did not interface with actors working behind the scenes. Based on this, T&S professionals using the Archive will (perhaps not surprisingly) note that it is rare for peer-reviewed inquiries to focus on the day-to-day, mostly invisible work required to sustain online platforms. Any insights for T&S practice should consider that the data is almost exclusively captured from users and content that is not removed.

We also found that the majority of investigations focused on highly contextualized datasets. Most academic researchers have access only to publicly accessible UGC. Due to data accessibility barriers, private content (a large portion of UGC) is not well represented in empirical scholarship. Additionally, researchers usually studied a particular community (DeCook et al. 2022), a particular social platform (most often Twitter) (Golbeck et al. 2017), and/or over a restricted time span (Myers West 2018). With the constantly evolving nature of T&S work, what was important for understanding harassment and cyberbullying at one moment in time in one subcommunity may not apply across different companies and teams.

Finally, a major challenge for deriving takeaways for professional practice from this corpus is that many articles center on users' self-reported experiences of harassment. The scale for user-reporting is often small, such as 20–30 interviews or several hundred survey responses. It is also unclear, based on what is represented in these manuscripts, how the populations studied and the analyses conducted by these researchers align with reports from platforms.

10 Takeaways for Researchers

For scholars interested in pursuing trust and safety scholarship, this review highlights several opportunities. First and foremost, the current body of peer-reviewed research centers on users. User perspectives are one of many in understanding online trust and safety. Current approaches to studying trust and safety issues are also largely limited to public-facing material. Seeking the input of T&S professionals and other stakeholders involved in platform governance would improve the quality and relevance of the research. Academic scholars and T&S professionals have different experiences and expertise. Combining them will increase the likelihood that research findings are relevant and applicable to trust and safety work.

The lack of professional perspectives was present in both the HCI and Communications scholarship. The HCI scholarship identified computational techniques to improve the accuracy of recognizing automated harassment/bullying content, reduce human moderation costs, and identify trends across users and contexts. The Communications

scholarship focused on elements of language and social behavior that predict producing undesirable content online.

However, based on what's reported in these investigations, T&S practitioners were not consulted in the design and testing of these interventions. In addition to these tools potentially informing T&S practice, deploying them in real-world contexts could serve as a critical source of data to confirm, challenge, and/or expand on the researchers' conclusions. Researchers are often testing technical specifications and moderation guidelines with highly curated datasets and crafted experimental scenarios. T&S professionals have a deep understanding of addressing real-time T&S issues at scale and what approaches may or may not be effective in their workflows. Additionally, T&S professionals face external pressure, such as government regulation, from stakeholders who do not fully understand T&S work. The perspectives of T&S professionals will help accurately contextualize data and conclusions in academic research. Thus, input and collaboration from T&S professionals would reveal potential issues with using these tools in T&S practice.

Additionally, we found that peer-reviewed scholarship does not make a meaningful distinction between platforms, the multifaceted stakeholder teams contributing to platform functionality including T&S teams, and individual T&S professionals. The form and function of platforms are complex. The approaches in this literature review did not capture the perspectives of the people and groups supporting online platforms, such as paid content moderators, regulators, and other teams who play a role in determining what material is flagged for review, what remains visible and accessible to users, and when and what is posted as official guidelines. What is visible to the public does not capture the full story of how and why decisions are made. Incorporating T&S professionals in the research process would offer more nuance from the involved parties contributing to the "output" of trust and safety activities on online platforms.

Next, there are major areas of contribution for trust and safety studies that have not been explored in peer-reviewed scholarship. In this review, we did not find consistent definitions of harassment and cyberbullying. Additionally, the definitions and theoretical approaches these investigations used were derived, understandably, from scholarship in the discipline to which the authors were contributing. Trust and safety scholarship exists, but it is fragmented across disciplines and terminologies. Future work on semi-systematic and systematic reviews would make valuable contributions to understanding the scope and content of trust and safety research. In tandem with recruiting T&S professionals as research participants, recruiting T&S professionals to have a seat at the table when developing approaches and takeaways would better contextualize the contributions and limitations of studies in T&S practice rather than somewhat idealized empirical settings.

There are ample opportunities to meaningfully collaborate with T&S professionals in peer-reviewed research. For example, subscribing to newsletters from industry voices and trust and safety nonprofits are potential avenues for identifying collaborators. The History of Trust & Safety Zotero Archive is one resource available to find academic scholars working on topics of interest (TSF 2023). The Trust and Safety Professional Association[3] hosts an annual conference, networking events, and other opportunities for connecting with T&S professionals around shared work and interests. Other resources include the Technology Coalition,[4] Tech Against Terrorism,[5] the Integrity Institute,[6] the

3. https://www.tspa.org/
4. https://www.technologycoalition.org/
5. https://techagainstterrorism.org/
6. https://integrityinstitute.org/

Digital Trust & Safety Partnership,[7] the Trust and Safety Foundation,[8] INHOPE,[9] and, of course, the Stanford Internet Observatory's annual Trust & Safety Research Conference. Inviting T&S professionals to the table, both as participants in research studies and as consultants and collaborators for research projects centered around trust and safety issues, will align research outcomes and contributions more closely with the scale and progression of the trust and safety challenges professionals face every day that impact millions of people worldwide.

7. https://dtspartnership.org/
8. https://trustandsafetyfoundation.org/
9. https://www.inhope.org/

References

Bhandari, Aparajita, Marie Ozanne, Natalya N. Bazarova, and Dominic DiFranzo. 2021. "Do You Care Who Flagged This Post? Effects of Moderator Visibility on Bystander Behavior." *Journal of Computer-Mediated Communication* 26, no. 5 (September): 284–300. https://doi.org/10.1093/jcmc/zmab007.

Blackwell, Lindsay, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. "Classification and Its Consequences for Online Harassment: Design Insights from HeartMob." *Proceedings of the ACM on Human-Computer Interaction* 1 (CSCW): 24:1–24:19. https://doi.org/10.1145/3134659.

Chadha, Kalyani, Linda Steiner, Jessica Vitak, and Zahra Ashktorab. 2020. "Women's Responses to Online Harassment." *International Journal of Communication* 14, no. 0 (January): 239–57. https://ijoc.org/index.php/ijoc/article/view/11683.

Chatzakou, Despoina, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. "Mean Birds: Detecting Aggression and Bullying on Twitter." arXiv: 1702.06877 [cs.CY].

DeCook, Julia R., Kelley Cotter, Shaheen Kanthawala, and Kali Foyle. 2022. "Safe from "harm": The governance of violence by platforms." *Policy & Internet* 14 (1): 63–78. https://doi.org/10.1002/poi3.290.

Dinakar, Karthik, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. "Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying." *ACM Transactions on Interactive Intelligent Systems* 2, no. 3 (September): 1–30. https://doi.org/10.1145/2362394.2362400.

Duguay, Stefanie, Jean Burgess, and Nicolas Suzor. 2020. "Queer women's experiences of patchwork platform governance on Tinder, Instagram, and Vine." *Convergence* 26, no. 2 (April): 237–52. https://doi.org/10.1177/1354856518781530.

Felmlee, Diane, Daniel DellaPosta, Paulina d. C. Inara Rodis, and Stephen A. Matthews. 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.

Feng, Chenlu, and Donghee Yvette Wohn. 2020. "Categorizing Online Harassment Interventions." In *2020 IEEE International Symposium on Technology and Society (ISTAS),* 255–65. November. https://doi.org/10.1109/ISTAS50296.2020.9462206.

Foley, Timothy, and Melda Gurakar. 2022. "Backlash or Bullying? Online Harassment, Social Sanction, and the Challenge of COVID-19 Misinformation." Number: 2, *Journal of Online Trust and Safety* 1 (February). https://doi.org/10.54501/jots.v1i2.31.

Founta, Antigoni-Maria, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. "Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior." arXiv: 1802.00393 [cs.SI].

Francisco, Sara C., and Diane H. Felmlee. 2022. "What Did You Call Me? An Analysis of Online Harassment Towards Black and Latinx Women." *Race and Social Problems* 14, no. 1 (March): 1–13. https://doi.org/10.1007/s12552-021-09330-7.

Golbeck, Jennifer, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, et al. 2017. "A Large Labeled Corpus for Online Harassment Research." In *Proceedings of the 2017 ACM on Web Science Conference,* 229–33. WebSci '17. New York, NY, USA: Association for Computing Machinery, June. https://doi.org/10.1145/3091478.3091509.

Goyal, Nitesh, Leslie Park, and Lucy Vasserman. 2022. "'You have to prove the threat is real': Understanding the needs of Female Journalists and Activists to Document and Report Online Harassment." In *CHI Conference on Human Factors in Computing Systems,* 1–17. New Orleans, LA: ACM, April. https://doi.org/10.1145/3491102.3517517.

Im, Jane, Sarita Schoenebeck, Marilyn Iriarte, Gabriel Grill, Daricia Wilkinson, Amna Batool, Rahaf Alharbi, et al. 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.

Jhaver, Shagun, Quan Ze Chen, Detlef Knauss, and Amy X. Zhang. 2022. "Designing Word Filter Tools for Creator-led Comment Moderation." In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems,* 205:1–205:21. CHI '22. New York: Association for Computing Machinery, April. https://doi.org/10.1145/3491102.3517505.

Jhaver, Shagun, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. "Online Harassment and Content Moderation: The Case of Blocklists." *ACM Transactions on Computer-Human Interaction* 25, no. 2 (March): 12:1–12:33. https://doi.org/10.1145/3185593.

Jones, Lisa, Kimberly Mitchell, and David Finkelhor. 2013. "Online Harassment in Context: Trends From Three Youth Internet Safety Surveys (2000, 2005, 2010)." *Psychology of Violence* 3 (January): 53–69. https://doi.org/10.1037/a0030309.

Klonick, Kate. 2023. "The End of the Golden Age of Tech Accountability." *The Klonickles: Newsletter of Kate Klonick, law professor and journalist* (March 3, 2023). https://klonick.substack.com/p/the-end-of-the-golden-age-of-tech.

Myers West, Sarah. 2018. "Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms." *New Media & Society* 20, no. 11 (November): 4366–83. https://doi.org/10.1177/1461444818773059.

Newton, Casey. 2023. "Have we reached peak trust and safety?" *Platformer* (June 9, 2023). https://www.platformer.news/p/have-we-reached-peak-trust-and-safety.

Nix, Naomi. 2023. "Meta to begin fresh layoffs, cutting heavily among business staff." *Washington Post* (May 23, 2023). https://www.washingtonpost.com/technology/2023/05/23/meta-layoffs-misinformation-facebook-instagram/.

Noble, Safiya Umoja, and Brendesha M. Tynes. 2016. *The Intersectional Internet: Race, Sex, Class, and Culture Online.* 2nd. Pieterlen, Berne, CHE: Peter Lang International Academic Publishers, February. https://doi.org/10.3726/978-1-4539-1717-6.

Page, Xinru, Bart P. Knijnenburg, Pamela Wisniewski, and Moses Namara. 2018. "Avoiding Online Harassment: The Socially Disenfranchised." In *Online Harassment,* edited by Jennifer Golbeck, 243–68. Human–Computer Interaction Series. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-78583-7_11.

Raisi, Elaheh, and Bert Huang. 2017. "Cyberbullying Detection with Weakly Supervised Machine Learning." In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017,* 409–16. ASONAM '17. New York, NY, USA: Association for Computing Machinery, July. https://doi.org/10.1145/3110025.3110049.

Razi, Afsaneh, Seunghyun Kim, Ashwaq Alsoubai, Gianluca Stringhini, Thamar Solorio, Munmun De Choudhury, and Pamela J. Wisniewski. 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.

Roose, Kevin, Casey Newton, Davis Land, Rachel Cohn, Jen Poyant, Alyssa Moxley, Dan Powell, et al. 2023. "Reddit Revolts, MrBeast's YouTube Empire and Peak Trust and Safety?" *The New York Times* (June). https://www.nytimes.com/2023/06/16/podcasts/reddit-revolts-mrbeasts-youtube-empire-and-peak-trust-and-safety.html.

Sanders, Teela, Gaynor Trueman, Kate Worthington, and Rachel Keighley. 2023. "Non-consensual sharing of images: Commercial content creators, sexual content creation platforms and the lack of protection." *New Media & Society* (May): 14614448231172711. https://doi.org/10.1177/14614448231172711.

Seering, Joseph, and Sanjay R. Kairam. 2022. "Who Moderates on Twitch and What Do They Do? Quantifying Practices in Community Moderation on Twitch." *Proceedings of the ACM on Human-Computer Interaction* 7 (GROUP): 18:1–18:18. https://doi.org/10.1145/3567568.

Suzor, Nicolas, Molly Dragiewicz, Bridget Harris, Rosalie Gillett, Jean Burgess, and Tess Van Geelen. 2019. "Human Rights by Design: The Responsibilities of Social Media Platforms to Address Gender-Based Violence Online." *Policy & Internet* 11, no. 1 (September 29, 2019): 84–103. https://doi.org/10.1002/poi3.185.

Trust & Safety Foundation. 2023. "History of Trust & Safety Zotero Library." Accessed January 19, 2024. https://www.zotero.org/groups/5077717/history_of_trust__safety/library.

Trust & Safety Professional Association. n.d. "Glossary: Common Terms & Definitions." Accessed January 19, 2024. https://www.tspa.org/curriculum/ts-curriculum/glossary/.

Vanian, Jonathan, and Hayden Field. 2023. "Tech layoffs ravage the teams that fight online misinformation and hate speech." *CNBC* (May 26, 2023). https://www.cnbc.com/2023/05/26/tech-companies-are-laying-off-their-ethics-and-safety-teams-.html.

Wohn, Donghee Yvette. 2019. "Volunteer Moderators in Twitch Micro Communities: How They Get Involved, the Roles They Play, and the Emotional Labor They Experience." In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems,* 1–13. Glasgow, UK: ACM, May. https://doi.org/10.1145/3290605.3300390.

Ybarra, Michele L., Kimberly J. Mitchell, Janis Wolak, and David Finkelhor. 2006. "Examining Characteristics and Associated Distress Related to Internet Harassment: Findings From the Second Youth Internet Safety Survey." *Pediatrics* 118, no. 4 (October): e1169–e1177. https://doi.org/10.1542/peds.2006-0815.

Zolides, Andrew. 2021. "Gender moderation and moderating gender: Sexual content policies in Twitch's community guidelines." *New Media & Society* 23, no. 10 (October): 2999–3015. https://doi.org/10.1177/1461444820942483.

Authors

Megan Knittel (knittel2@msu.edu) is an Assistant Professor in the Department of Media & Information at Michigan State University, where she also received her PhD in information and media. She is a Research Assistant with the Trust & Safety Foundation's History of Trust & Safety project. Megan is most interested in understanding how social technology changes the way we relate to ourselves and others, and in supporting the development of safe and beneficial online platforms.

Amanda Menking (amanda@trustandsafetyfoundation.org) joined the Trust & Safety Professional Association (TSPA) and the Trust & Safety Foundation (TSF) after spending a decade in academia, researching bias, knowledge production, and safety in online communities and teaching courses about design, research, and human-computer interaction. Amanda is most interested in doing meaningful work with thoughtful, kind people with the aim of building more just and equitable worlds.

Acknowledgements

We would like to acknowledge the invited subject matter experts who provided their knowledge and recommendations for developing the History of Trust and Safety Project Archive. In no particular order: Robyn Caplan, Oliver Haimson, Nic Suzor, Lucinda Nelson, Jeff Lazarus, Leonie Tanczer, Joseph Seering, Ysabel Gerrard, and Mitali Thakor. We also would like to thank Sarah Godlewski and Charlotte Willner for reviewing this manuscript.

Keywords

Trust and safety; literature review; harassment; cyberbullying; Human Computer Interaction; Communications.