

COMPASS: The 2nd International Workshop on Socio-technical Approaches to Content Moderation and Platform Governance

Robyn Caplan
Duke University
Durham, North Carolina, USA
robyn.caplan@duke.edu

Stefano Cresci
IIT-CNR
Pisa, Italy
stefano.cresci@iit.cnr.it

Catalina Goanta
Faculty of Law, Governance and
Economics
Utrecht University
Utrecht, Netherlands
e.c.goanta@uu.nl

Shagun Jhaver
Rutgers University
New Brunswick, New Jersey, USA
shagun.jhaver@rutgers.edu

Sanjay R Kairam
OpenAI
San Francisco, California, USA
sanjay.kairam@gmail.com

Koustuv Saha
Siebel School of Computing and Data
Science
University of Illinois
Urbana-Champaign
Urbana, Illinois, USA
ksaha2@illinois.edu

Savvas Zannettou
TU Delft
Delft, Netherlands
s.zannettou@tudelft.nl

Abstract

Content moderation is an essential yet increasingly complex component of maintaining safe and inclusive online environments. Platforms navigate a growing range of harms, including misinformation, hate speech, harassment, exploitative material, and emerging challenges associated with generative AI and evolving platform infrastructures. At the same time, content moderation extends well beyond automated detection: it also involves defining community guidelines, designing intervention strategies, assessing downstream effects, promoting desirable contributions, and accounting for cultural, legal, and organizational differences across contexts. Addressing these challenges requires a multidisciplinary perspective that connects socio-technical approaches with human-centered design, governance, ethics, and regulation. COMPASS will thus foster a dialogue on content moderation and platform governance among researchers and practitioners from CSCW, HCI, computational social science, law, communication, and related fields. Building on the strong engagement of the first edition, successfully held at AAAI ICWSM 2025, this workshop will provide a venue to examine moderation as a socio-technical problem shaped by platforms, policies, communities, and institutional constraints. By bringing together diverse perspectives and emphasizing interaction and exchange, COMPASS will strengthen cross-disciplinary collaboration and advance timely, human-centered approaches to online safety and governance.

CCS Concepts

• **Human-centered computing** → **Collaborative and social computing**; • **Information systems** → *Social networks*; • **Applied computing** → Law, social and behavioral sciences.

Keywords

Content Moderation, Platform Governance, Socio-technical Systems, Online Communities, Trust and Safety, Human-centered AI

ACM Reference Format:

Robyn Caplan, Stefano Cresci, Catalina Goanta, Shagun Jhaver, Sanjay R Kairam, Koustuv Saha, and Savvas Zannettou. 2026. COMPASS: The 2nd International Workshop on Socio-technical Approaches to Content Moderation and Platform Governance. In *Companion of the Computer-Supported Cooperative Work and Social Computing (CSCW Companion '26)*, October 10–14, 2026, Salt Lake City, UT, USA. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3785651.3818816>

1 Scientific Focus and Topics

Content moderation is a critical and enduring component of online platforms, serving as a frontline response to a wide range of harms emerging from user-generated content and behavior [11, 13]. These include mis- and disinformation [6, 26], hate speech [2, 5, 14], and exploitative or abusive material [40], as well as emerging challenges associated with generative AI, evolving platform infrastructures, and social commerce. As platforms expand in scope and societal impact, moderation must also address shifting legal and regulatory requirements, including consumer protection and transparency obligations. Content moderation is therefore not only a technical challenge, but also a social and governance problem requiring continuous adaptation.



This work is licensed under a Creative Commons Attribution 4.0 International License. *CSCW Companion '26, Salt Lake City, UT, USA*
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2378-0/26/10
<https://doi.org/10.1145/3785651.3818816>

Effective moderation involves multiple interconnected components. Platforms define community guidelines that establish norms and expectations [30], deploy computational methods to detect policy violations [10, 28, 42], and implement interventions such as content removals, frictions, labeling, visibility reductions, norm promotions, or user sanctions [1, 15, 16, 18, 25, 27, 34, 41]. Beyond platform-level policy, individual communities and their moderators constitute an additional layer of governance, setting community goals, defining local norms, and customizing interventions to meet community-specific requirements [8, 21, 33]. These processes involve both automated systems and human decision-making, and their outcomes must be evaluated across multiple levels, including individual wellbeing, community health and flourishing, platform health, and broader societal impacts [22, 38]. These components introduce numerous multidisciplinary challenges spanning methodological, social, normative, and ethical domains. Methodologically, ensuring data access [3, 12], as well as robustness, scalability, and causal validity, remains difficult, particularly when evaluating the downstream effects of moderation interventions [17, 34]. Socially, moderation systems must account for human factors, including user perceptions, cultural differences, and the psychological impact on both moderators and affected users [4, 7, 31]. Normatively, increasing regulatory pressure (e.g., the EU Digital Services Act and AI Act) raises requirements for transparency, accountability, and fairness [23, 36, 37], while also highlighting tensions between harm prevention and freedom of expression [35]. At the same time, emerging paradigms such as user-driven or decentralized moderation introduce new forms of agency alongside risks such as increased cognitive burden and fragmentation [19, 20]. In parallel, the rapid diffusion of generative AI is reshaping the moderation landscape by both amplifying risks, such as through the scalable production of synthetic, misleading, or harmful content [39], and creating new opportunities as the use of simulation environments and AI-driven agents to study user behavior and evaluate the potential effects of moderation strategies before deployment [9, 32]. Ethically, platforms must balance competing values including safety, privacy, and autonomy [24]. Despite growing interest in these challenges, research has often focused narrowly on detection, overlooking broader questions of governance, intervention design, and evaluation. Recent industry shifts, such as changes in moderation practices and resource allocation [29], further illustrate the evolving and contested nature of moderation systems, shaped by technical, political, economic, and societal forces.

Building on the success of the first COMPASS workshop,¹ which brought together a large interdisciplinary community and received strong engagement, this second edition continues to advance a holistic, socio-technical perspective on content moderation. The original organizing team is maintained and further expanded for this second edition, combining expertise across computational and socio-technical methods, social computing, HCI, platform governance, and industry practices, and reinforcing the workshop's role as a bridge across research communities. In line with CSCW's focus on human-centered and collaborative systems, the workshop will encourage submissions and discussions on topics including (but not limited to):

- *Community guidelines and governance*: Design, evolution, and impact of platform policies, and their alignment with community norms and regulatory frameworks.
- *Moderation interventions and outcomes*: Design and evaluation of interventions (e.g., removals, warnings, visibility modifications), including causal inference and downstream effects.
- *Metrics and evaluation*: Measures of moderation effectiveness, platform health, and societal impact.
- *Human-centered moderation*: Interactions between human moderators, users, and automated systems; psychological and experiential dimensions.
- *Cultural and societal dimensions*: Cross-cultural differences, inclusivity, and the role of diverse communities in shaping moderation practices.
- *Legal and regulatory implications*: Impacts of frameworks such as the DSA and AI Act on platform governance and accountability.
- *Ethics and values*: Trade-offs between safety, privacy, autonomy, and freedom of expression.
- *Data and transparency*: Access to moderation data, auditing practices, and reproducibility challenges.
- *Decentralized and participatory moderation*: Community-driven and distributed approaches to moderation and governance.
- *Emerging technologies*: Moderation challenges related to generative AI, LLM-powered moderation interventions, live and real-time content, and new platform modalities.
- *Special populations*: Unique moderation concerns in the context of special groups, such as politicians, activists, journalists, and children.

2 Format and Schedule

We propose a full-day, highly interactive workshop building on the successful format of the first COMPASS edition. The workshop will combine structured presentations with participatory activities designed to foster discussion, collaboration, and community-building. The program includes a keynote, two paper presentation sessions, a thematic panel, and a collaborative brainstorming session. All sessions will incorporate moderated discussions, audience participation, and opportunities for cross-disciplinary exchange. Paper sessions will prioritize shorter presentations followed by extended discussion. Paper session #2 (featuring *resource*, *vision*, and *work-in-progress* papers) is specifically designed to stimulate discussion and seed the final collaborative session, which specifically focuses on collectively identifying open challenges, research directions, and opportunities for collaboration. COMPASS will also provide dedicated opportunities for informal interaction and networking throughout the day.

The tentative schedule is as follows (subject to adjustment based on submissions and speaker availability):

9:00—Welcome and introduction.

9:30—Keynote.

10:30—Coffee break.

10:45—Paper session #1 (*Research papers or posters*).

12:30—Lunch break.

13:30—Panel discussion.

¹<https://sites.google.com/view/workshop-compass25/home>

14:30—Paper session #2 (*Resource, vision, and work-in-progress papers*).

15:45—Coffee break.

16:00—Collaborative brainstorming and breakout discussions.

17:00—Synthesis of discussion outcomes.

17:15—Closing remarks.

We will accommodate up to 25–30 participants to ensure a high level of interaction and meaningful discussion, consistent with the format successfully tested in the previous edition. We require a standard conference room setup, including a projector, screen, microphones (depending on room size), and reliable internet connectivity. We will solicit the accessibility needs of our participants in advance of the workshop and work closely with the Accessibility Chairs to fulfill those needs as best as possible. We expect that participants will develop a nuanced understanding of the state of the art in content moderation research, including current approaches, key debates, and open problems, opportunities for collaboration, and promising directions for future scholarship.

3 Submissions and TPC

We encourage four types of submissions: (i) *research papers*, (ii) *resource papers* (e.g., demo, dataset, software), (iii) *vision and position papers*, and (iv) *work-in-progress papers*. This range of formats is intended to support contributions at different stages of maturity and from diverse methodological and disciplinary perspectives. Each submission will be reviewed by three anonymous reviewers from our experienced and multidisciplinary technical program committee (TPC). Submissions will be evaluated based on their contribution, novelty, rigor, and relevance to the workshop topics, as well as their potential to stimulate discussion and cross-disciplinary exchange. The diversity of the organizing team and TPC (both geographic and scientific) ensures broad coverage across relevant fields, including HCI, CSCW, computational social science, and platform governance. We are also committed to fostering an inclusive and constructive review process that values diverse perspectives, research approaches, and scholarly backgrounds.

4 Other Related Workshops

Table 1 presents a selection of recent workshops related to content moderation and online harms. While these venues have contributed valuable insights, they tend to be situated primarily within AI, NLP, or data mining communities, and often emphasize computational approaches to detecting harmful content (e.g., CySoc, WOAAH, MMCM, Integrity). Other workshops (e.g., RegNLP, NLLP) focus on the analysis of legal or policy texts, but typically without engaging with the full lifecycle of content moderation in practice.

As a result, existing workshops only partially capture the inherently socio-technical nature of content moderation, where automated systems, human decision-making, platform governance, user communities, and regulatory frameworks are deeply intertwined. There is currently no dedicated venue that brings together these perspectives in a balanced and integrated way, nor one that systematically bridges computational approaches with human-centered, organizational, and policy-oriented analyses. COMPASS addresses this gap by explicitly positioning content moderation as a multidisciplinary and socio-technical problem. Building on the strong

	related workshop	venue
CySoc	<i>Workshop on Cyber Social Threats</i>	ICWSM
Integrity	<i>Workshop on Integrity in Social Networks and Media</i>	WSDM
WOAH	<i>Workshop on Online Abuse and Harms</i>	ACL
MMCM	<i>Workshop on Multimodal Content Moderation</i>	CVPR
NLLP	<i>Workshop on Natural Legal Language Processing</i>	EMNLP
RegNLP	<i>Regulatory Natural Language Processing Workshop</i>	COLING

Table 1: Related workshops.

engagement of the first edition, we will create a space where researchers and practitioners can examine moderation not only as a technical challenge, but also as a question of design, governance, and societal impact. This includes topics such as the development and interpretation of community guidelines, the design and evaluation of interventions, the role of human moderators and users, and the implications of emerging regulatory frameworks. This positioning makes COMPASS particularly well aligned with CSCW, a venue that foregrounds the study of complex socio-technical systems and the interaction between computational technologies and social processes.

5 Organizers

COMPASS is organized by a multidisciplinary team of established scholars and industry experts with expertise spanning HCI, social computing, computational social science, law, and platform governance. This breadth of expertise is central to the workshop’s goal of advancing content moderation as a socio-technical and governance challenge, and ensures the ability to engage a broad, diverse, and international research community. Building on the successful organization of the first COMPASS workshop at ICWSM 2025,² the team brings continuity, complementary perspectives, and demonstrated success and experience in fostering interdisciplinary dialogue on content moderation.

COMPASS will be run on-site by Shagun Jhaver, Sanjay Kairam, and Koustuv Saha, who are established members of the CSCW community and have prior experience organizing workshops and panels at CSCW and related venues. They will attend CSCW’26 in person and lead all COMPASS in-person activities, including session facilitation and coordination. The rest of the organizing team, comprising Robyn Caplan, Stefano Cresci, Catalina Goanta, and Savvas Zannettou, have contributed to shaping the workshop’s vision and this proposal, and will actively support in drafting the call for papers, managing the review process, advertising the workshop, and structuring the program. This distributed organizing model, successfully adopted in the previous edition, enables sustained collaboration across research communities and geographical contexts.

- **Robyn Caplan** is an Assistant Professor at Duke University’s Sanford School of Public Policy and a founding member of the Platform Governance Research Network. She is a leading expert in *platform governance* and *media policy*, with a focus on how organizational structures and policy

²<https://www.icwsm.org/2025/>

decisions influence content moderation. Her research critically examines transparency, accountability, and the role of governance in shaping online speech. She has extensive experience organizing interdisciplinary events and workshops across academia and policy communities.

- **Stefano Cresci** is a Senior Researcher at the Institute for Informatics and Telematics of the National Research Council (IIT-CNR) in Italy. Stefano's scientific interests broadly lie at the intersection of *Web*, *network*, and *data science*, with a specific focus on content moderation, coordinated online behavior, and human-centered AI. On these topics, he published extensively in venues such as PNAS, WWW, ICWSM, CSCW, and WebSci. Currently, he leads a prestigious ERC project on data-driven and personalized content moderation. Previously, he received multiple awards including the IEEE Next-Generation Data Scientist Award and the ERCIM Cor Baayen Young Researcher Award. He recently co-organized successful events at AAAI ICWSM, ACM WSDM, IEEE DSAA, and SocInfo.
- **Catalina Goanta** is an Associate Professor in Private Law and Technology at the University of Utrecht. She is an esteemed scholar in the area of *platform governance*, with multiple notable contributions on understanding the impact of content monetization on social media and on reinterpreting private law fairness in the context of online platforms. Catalina's expertise at the intersection of computational science and law provides critical insights into legal frameworks and helps in navigating the complex interplay between regulatory requirements, platform accountability, and user rights. She co-organized multiple scientific events, including a workshop series at ACL EMNLP.
- **Shagun Jhaver** (*Local Organizer*) is an Assistant Professor in the School of Communication & Information at Rutgers University. Shagun is an expert in *human-computer interaction* and a leading researcher in the CSCW and CHI communities. His research focuses on improving content moderation processes on digital platforms from a user-centered perspective. He examines how platform design, policies, and technical affordances shape user experiences and moderation outcomes, contributing empirically grounded insights into fairness, transparency, and online safety. He has extensive experience organizing workshops and panels at CSCW.
- **Sanjay Kairam** (*Local Organizer*) is a Member of Technical Staff at OpenAI, where he works on evaluation for frontier AI systems. His research and applied work has focused on the measurement and evaluation of socio-technical systems, including online communities, content platforms, and AI systems. In previous roles at Reddit and Twitch, he has developed methods to assess community health, user experience, and the impact of platform design on social interactions. His work combines *computational social science*, *causal inference*, *experimentation*, and *machine learning* to study how platforms shape user behavior and well-being. His industry experience provides a valuable practitioner-oriented perspective on the challenges of content moderation and platform governance.
- **Koustuv Saha** (*Local Organizer*) is an Assistant Professor of Computer Science at the University of Illinois Urbana-Champaign. His research examines how online and AI technologies shape and reveal human behaviors and wellbeing, including in situated contexts such as college campuses and workplaces. He uses *machine learning*, *natural language*, and *causal inference* methods to study human behavior and wellbeing using digital data, including social media and multimodal sensing data. His work further explores content moderation, trust and safety, and the wellbeing effects of online content exposure and engagement, focusing on how AI- and data-driven systems govern online content, shape user experiences, and influence individuals and communities.
- **Savvas Zannettou** is an Assistant Professor at TU Delft and an associated researcher with the Max Planck Institute for Informatics. He is well-known in the ICWSM, WWW, IMC, and CSCW communities, thanks to his multiple contributions in the area of *computational social science*. He has extensive research experience in topics related to the proposed workshop, including understanding and detecting hate speech, analyzing the efficacy of content moderation interventions, auditing content moderation interventions on platforms like TikTok and Twitter/X, as well as analyzing regulatory and transparency requirements in online spaces.

Acknowledgments

This work is partly supported by the European Union with the ERC project DEDUCE under grant #101113826.

References

- [1] Shiza Ali, Mohammad Hammas Saeed, Esraa Aldreabi, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. 2021. Understanding the effect of deplatforming on social networks. In *The 13th ACM Web Science Conference (WebSci)*.
- [2] Helena Bonaldi, Yi-Ling Chung, Gavin Abercrombie, and Marco Guerini. 2024. NLP for Counterspeech against Hate: A Survey and How-To Guide. In *Findings of the Association for Computational Linguistics (NAACL)*.
- [3] Aldo Cerulli, Lorenzo Cima, Benedetta Tessa, Serena Tardelli, and Stefano Cresci. 2026. The Big Ban Theory: A pre- and post-intervention dataset of online content moderation actions. In *The 20th International AAAI Conference on Web and Social Media (ICWSM)*. AAAI.
- [4] Aldo Cerulli, Benedetta Tessa, Giuseppe La Selva, Oronzo Mazzeo, Lorenzo Cima, Lucia Monacis, and Stefano Cresci. 2025. Dark Personality Traits and Online Toxicity: Linking Self-Reports to Reddit Activity. *arXiv:2512.10113* (2025).
- [5] Lorenzo Cima, Alessio Miaschi, Amaury Trujillo, Marco Avvenuti, Felice Dell'Orletta, and Stefano Cresci. 2025. Contextualized counterspeech: Strategies for adaptation, personalization, and evaluation. In *The 34th ACM Web Conference (WWW)*.
- [6] Thomas H Costello, Gordon Pennycook, and David G Rand. 2024. Durably reducing conspiracy beliefs through dialogues with AI. *Science* 385, 6714 (2024).
- [7] Stefano Cresci, Amaury Trujillo, and Tiziano Fagnì. 2022. Personalized interventions for online moderation. In *The 33rd ACM Conference on Hypertext and Social Media (HT)*.
- [8] Amanda LL Cullen and Sanjay R Kairam. 2022. Practicing moderation: Community moderation as reflective practice. *Proceedings of the ACM on Human-computer Interaction* 6, CSCW1 (2022), 1–32.
- [9] Giacomo Fidone, Lucia Passaro, and Riccardo Guidotti. 2026. Evaluating Online Moderation via LLM-Powered Counterfactual Simulations. In *The 2026 AAAI Conference on Artificial Intelligence (AAAI)*.
- [10] Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)* 51, 4 (2018).
- [11] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press.
- [12] Catalina Goanta, Savvas Zannettou, Rishabh Kaushal, Jacob van de Kerkhof, Thales Bertaglia, Taylor Annabell, Haoyang Gui, Gerasimos Spanakis, and Adriana Iamnitchi. 2025. The great data standoff: Researchers vs. platforms under

- the Digital Services Act. In *The 20th International AAAI Conference on Web and Social Media (ICWSM)*.
- [13] James Grimmelmann. 2015. The virtues of moderation. *Yale Journal of Law and Technology* 17 (2015), 42.
 - [14] Manoel Horta Ribeiro, Pedro Calais, Yuri Santos, Virgílio Almeida, and Wagner Meira Jr. 2018. Characterizing and detecting hateful users on Twitter. In *The 12th International AAAI Conference on Web and Social Media (ICWSM)*.
 - [15] Manoel Horta Ribeiro, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Gianluca Stringhini, Emiliano De Cristofaro, and Robert West. 2021. Do platform migrations compromise content moderation? Evidence from r/The_Donald and r/Incels. In *The 24th ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW)*.
 - [16] Manoel Horta Ribeiro, Robert West, Ryan Lewis, and Sanjay Kairam. 2025. Post guidance for online communities. *Proceedings of the ACM on Human-Computer Interaction* 9, 2 (2025), 1–26.
 - [17] Homa Hosseinmardi, Amir Ghasemian, Miguel Rivera-Lanas, Manoel Horta Ribeiro, Robert West, and Duncan J Watts. 2024. Causally estimating the effect of YouTube's recommender system using counterfactual bots. *Proceedings of the National Academy of Sciences* 121, 8 (2024).
 - [18] Shagun Jhaver. 2025. Bans vs. Warning Labels: Examining Bystanders' Support for Community-wide Moderation Interventions. *Accepted in ACM Trans. Comput.-Hum. Interact.* (2025).
 - [19] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Trans. Comput.-Hum. Interact.* 26, 5 (2019).
 - [20] Shagun Jhaver, Alice Qian Zhang, Quan Ze Chan, Nikhila Natarajan, Ruotong Wang, and Amy X. Zhang. 2023. Personalizing Content Moderation on Social Media: User Perspectives on Moderation Choices, Interface Design, and Labor. *Proc. ACM Hum.-Comput. Interact.* (2023), 33 pages.
 - [21] Sanjay R Kairam and Jeremy Foote. 2024. How founder motivations, goals, and actions influence early trajectories of online communities. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–11.
 - [22] Sanjay R Kairam, Melissa C Mercado, and Steven A Sumner. 2022. A social-ecological approach to modeling sense of virtual community (SOVC) in livestreaming communities. *Proceedings of the ACM on human-computer interaction* 6, CSCW2 (2022), 1–35.
 - [23] Rishabh Kaushal, Jacob Van De Kerkhof, Catalina Goanta, Gerasimos Spanakis, and Adriana Iamnitchi. 2024. Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.
 - [24] Anastasia Kozyreva, Stefan M Herzog, Stephan Lewandowsky, Ralph Hertwig, Philipp Lorenz-Spreen, Mark Leiser, and Jason Reifler. 2023. Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences* 120, 7 (2023).
 - [25] Charlotte Lambert, Frederick Choi, and Eshwar Chandrasekharan. 2024. "Positive reinforcement helps breed positive behavior": Moderator Perspectives on Encouraging Desirable Behavior. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2 (2024).
 - [26] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018).
 - [27] Paddy Leerssen. 2023. An end to shadow banning? Transparency rights in the Digital Services Act between content moderation and curation. *Comput. Law Secur. Rep.* 48, 105790 (April 2023), 105790.
 - [28] Lorenzo Mannocci, Michele Mazza, Anna Monreale, Maurizio Tesconi, and Stefano Cresci. 2024. Detection and characterization of coordinated online behavior: A survey. *arXiv:2408.01257* (2024).
 - [29] Meta Platforms, Inc. 2025. More Speech, Fewer Mistakes: Meta's Approach to Evolving Content Moderation. <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>. Accessed: January 13, 2025.
 - [30] Mohit Singhal, Chen Ling, Pujan Paudel, Poojitha Thota, Nihal Kumarswamy, Gianluca Stringhini, and Shirin Nilizadeh. 2023. SoK: Content moderation in social media, from guidelines to enforcement, and research to practice. In *The 8th IEEE European Symposium on Security and Privacy (EuroS&P)*.
 - [31] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease. 2021. The psychological well-being of content moderators: The emotional labor of commercial moderation and avenues for improving support. In *The 39th CHI Conference on Human Factors in Computing Systems (CHI)*.
 - [32] Erika Elizabeth Taday Morocho, Lorenzo Cima, Tiziano Fagni, Marco Avvenuti, and Stefano Cresci. 2026. Assessing the reliability of persona-conditioned LLMs as synthetic survey respondents. In *The 35th ACM Web Conference (WWW Workshops)*. ACM.
 - [33] Benedetta Tessa, Lorenzo Cima, Amaury Trujillo, Marco Avvenuti, and Stefano Cresci. 2025. Beyond trial-and-error: Predicting user abandonment after a moderation intervention. *Engineering Applications of Artificial Intelligence* 162 (2025).
 - [34] Amaury Trujillo and Stefano Cresci. 2022. Make Reddit Great Again: Assessing community effects of moderation interventions on r/The_Donald. In *The 25th ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW)*.
 - [35] Amaury Trujillo and Stefano Cresci. 2023. One of many: Assessing user-level effects of moderation interventions on r/The_Donald. In *The 15th International ACM Web Science Conference (WebSci)*.
 - [36] Amaury Trujillo, Tiziano Fagni, and Stefano Cresci. 2025. The DSA Transparency Database: Auditing self-reported moderation actions by social media. In *The 28th ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW)*.
 - [37] Amaury Trujillo, Benedetta Tessa, and Stefano Cresci. 2026. Disarranged harmonization of transparency reporting by social media platforms under the Digital Services Act. *arXiv:2605.17655* (2026).
 - [38] Galen Weld, Carl Pearson, Bradley Spahn, Tim Althoff, Amy X. Zhang, and Sanjay Kairam. 2026. How Conversational Structure and Style Shape Online Community Experiences. In *The 20th International AAAI Conference on Web and Social Media (ICWSM)*.
 - [39] Kai-Cheng Yang and Filippo Menczer. 2024. Anatomy of an AI-powered malicious social botnet. *Journal of Quantitative Description: Digital Media* 4 (2024).
 - [40] Greyson K Young. 2022. How much is too much: The difficulties of social media content moderation. *Information & Communications Technology Law* 31, 1 (2022).
 - [41] Savvas Zannettou. 2021. "I Won the Election!": An empirical analysis of soft moderation interventions on Twitter. In *The 15th International AAAI Conference on Web and Social Media (ICWSM)*.
 - [42] Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)* 53, 5 (2020).