

COMPASS: 1st International Workshop on Computational Approaches to Content Moderation and Platform Governance

Stefano Cresci^{1*}, Savvas Zannettou², Catalina Goanta³, Shagun Jhaver⁴, Robyn Caplan⁵

¹IIT-CNR, Italy ²TU Delft, Netherlands ³University of Utrecht, Netherlands

⁴Rutgers University, NJ, USA ⁵Duke University, NC, USA

stefano.cresci@iit.cnr.it, s.zannettou@tudelft.nl, e.c.goanta@uu.nl, sj917@comminfo.rutgers.edu, robyn.caplan@duke.edu

Abstract

Content moderation is an essential—yet increasingly complex—component of maintaining safe and inclusive online environments. Nowadays, platforms navigate a growing range of harms including misinformation, hate speech, harassment, and exploitative material. The rapid advancement of technologies such as generative AI, metaverses, or recommender systems for online content further complicates this landscape, introducing new forms of misbehavior and challenging the adaptability of existing strategies. Moreover, with the rise of social commerce, social media platforms need to moderate user activities around acts of commerce according to completely different legal frameworks than freedom of expression. While significant research has focused on developing detection mechanisms for multiple online harms, other critical aspects of the moderation process remain underexplored, including the definition of community guidelines, the design of intervention strategies, and the development of robust methodologies to evaluate moderation outcomes. Addressing these challenges requires a holistic, multidisciplinary approach that bridges computational methods, human factors, ethical considerations, and evolving transnational regulatory frameworks. This workshop aims to foster a comprehensive dialogue on content moderation, engaging researchers and practitioners from diverse backgrounds, including social computing and computational social science, law, human-computer interaction, and platform governance. By bringing together varied perspectives, the workshop will advance understanding, spark collaboration, and drive innovation in tackling the full spectrum of content moderation tasks. ICWSM, with its interdisciplinary focus on computational and social sciences, is the ideal venue for starting this workshop series, creating a platform for meaningful discussion and actionable insights into the pressing content moderation challenges that shape today’s digital ecosystems.

Scientific Focus and Topics

Content moderation is a critical and enduring component of online platforms, as it represents the frontline defense against the many harms that arise from user-generated content and behavior (Gillespie 2018). The spread of mis- and disinformation (Lazer et al. 2018; Costello, Pennycook, and Rand 2024), the propagation of hate speech (Horta Ribeiro

et al. 2018; Bonaldi et al. 2024), and the distribution of non-consensual pornography or the sexual exploitation of minors (Young 2022) are but a few notable examples of online harms that require accurate and timely moderation (Saltz et al. 2021). To this more traditional understanding of content moderation we can add emerging issues around social commerce, including native advertising such as influencer marketing and live selling, which bring with them consumer protection standards focused on the conspicuity of advertising and shopping through disclosures and other relevant information. Content moderation is—and is bound to remain—instrumental in ensuring safe, trustworthy, transparent, and inclusive online spaces. The rapid technological advancement that characterizes current times introduces frequent disruptive innovations, such as generative AI and metaverses. Consequently, the landscape of online threats and misbehavior, as well as subsequent issues of legal compliance, is bound to evolve continuously, demanding that platforms remain vigilant and adaptable in their moderation strategies.

Effective content moderation encompasses several key steps. Firstly, platforms establish community guidelines to delineate acceptable content and behaviors (Singhal et al. 2023). These guidelines are the foundation upon which moderation efforts are built. Subsequently, detection methods—often leveraging advanced computational techniques—are employed to identify content and behaviors that breach the guidelines (Fortuna and Nunes 2018; Zhou and Zafarani 2020; Cresci 2020). Once problematic content is detected, administrators, moderators, and algorithms engage in decision-making processes to determine the appropriate response. This may involve removing content (Horta Ribeiro et al. 2021; Ali et al. 2021; Trujillo and Cresci 2022), issuing warnings (Jhaver 2025; Zannettou 2021), reducing visibility (Leerssen 2023), or implementing other interventions designed to mitigate the resulting harms (Cox et al. 2016). The effectiveness of the interventions is not a given, and it must be continually assessed through key performance indicators (KPIs) that monitor the overall health of the platform and the efficacy of the moderation process as a whole.

These steps introduce a multitude of multidisciplinary challenges spanning methodological, social, normative, and ethical domains. Methodologically, ensuring the accuracy and scalability of detection algorithms is a complex task,

*Main contact

particularly when dealing with the vast and varied nature of online content. Additionally, obtaining reliable causal estimates of the effects of moderation interventions presents significant challenges, as it requires robust experimental or quasi-experimental designs to accurately measure their impact on user behavior and platform health (Hosseinmardi et al. 2024; Trujillo and Cresci 2022). Normatively, there is a pressing need for platform transparency and fairness in content moderation practices, as emphasized by recent laws such as the EU Digital Services Act and the AI Act, which set rigorous standards for accountability and transparency (Kaushal et al. 2024; Trujillo, Fagni, and Cresci 2025). Researchers have also highlighted the need to provide moderated users with comprehensive information about the reasons behind content removals (Jhaver, Bruckman, and Gilbert 2019; Jhaver et al. 2019a). Socially, it is crucial to consider both algorithmic and human factors in moderation decisions to avoid unintended consequences (Russo et al. 2023; Cima et al. 2024), such as the suppression of free speech or the marginalization of certain groups (Trujillo and Cresci 2023). Moreover, the multifaceted and nuanced nature of human behavior adds layers of complexity to the analyses, making it essential to account for diverse user perspectives (Cresci, Trujillo, and Fagni 2022), cultural contexts, and psychological impacts on both the moderators and the moderated (Steiger et al. 2021). Beyond top-down moderation mechanisms, tools that enable end-users (Jhaver et al. 2023) or local community administrators (Jhaver et al. 2019b) to personalize content feeds offer greater control but raise concerns about the required cognitive labor and filter bubbles. Ethically, the responsibility of platforms to balance harm prevention with respect for privacy and user autonomy presents additional challenges (Kozyreva et al. 2023). Further, the regulation of politicians' speech raises debates about preventing harm versus respecting constituents' right to know the views of their elected officials.

In summary, while the adoption of quantitative and computational approaches has become instrumental in addressing these issues, most computational research to date focuses narrowly on developing detection mechanisms, often neglecting the broader aspects of the moderation process. Meta's recent push to reduce content moderation (Meta Platforms, Inc. 2025) underscores the intricate balance between free expression, platform responsibility, dedicated resources, regulatory compliance, and technical feasibility. These changes, driven by competing interests in politics, ethics, law, and technology, exemplify the dynamic landscape of content moderation and the challenges of designing adaptable, transparent, and effective moderation systems. Addressing these challenges necessitates a plurality of approaches and expertise, incorporating multiple viewpoints and diverse methodologies. The multidisciplinary nature of ICWSM makes it the ideal venue for discussing computational approaches to content moderation. The inaugural edition of this timely workshop series thus aims to provide a computational window into the expanding multidisciplinary challenges of content moderation, fostering a holistic and integrated discourse. The workshop will encourage submissions and discussion on and beyond the following topics:

- *Community guidelines and policy formulation*: Computational studies of the creation, adaptation, and impact of community guidelines and terms of service, and their alignment with evolving online behaviors and threats.
- *Intervention strategies and efficacy*: Development of new moderation interventions and computational investigations on various interventions (e.g., content removal, user warnings, temporary bans) and their effectiveness. Methodological approaches to obtain reliable causal estimates of moderation effects.
- *Measurement and metrics for moderation success*: Development and validation of key performance indicators (KPIs) to assess the health of online platforms and the success of content moderation efforts.
- *Human factors in content moderation*: Computational examinations of the role and impact of human factors in content moderation, including the psychological effects of moderation on both moderators and moderated users, and the interaction between human and algorithmic moderation.
- *Cross-cultural and social dimensions of content moderation*: Computational solutions to the challenges posed by diverse cultural contexts and user interactions, ensuring inclusive and equitable moderation practices.
- *Normative and legal challenges in content moderation*: Computational analyses of the impact of recent regulations such as the EU Digital Services Act (DSA) and AI Act on platform transparency, accountability, and fairness in content moderation practices.
- *Ethical considerations in content moderation*: Computational approaches to balance harm prevention with respect for user privacy and autonomy, and addressing ethical dilemmas in the implementation of moderation policies.
- *Data access for content moderation*: Innovative methods to obtain reliable and comprehensive datasets on removed content and content moderation decisions. Access to such reliable and comprehensive datasets is paramount for assessing the effectiveness of existing moderation interventions and for auditing online social platforms.
- *Decentralized content moderation*: Computational approaches to analyze decentralized content moderation in emerging decentralized communities and design of effective moderation interventions in spaces where moderation processes are distributed.
- *Future technologies and content moderation*: Anticipating the challenges and opportunities presented by emerging technologies such as generative AI and metaverses in the context of content moderation. Investigating emerging challenges in environments requiring real-time content moderation, like live streams or other fast-moving content.

Format and Schedule

We propose a full-day workshop beginning with a meet and greet session and consisting of diversified activities, including a keynote talk, two paper presentation sessions, a the-

Reuben Binns, *University of Oxford*, UK
 Eshwar Chandrasekharan, *University of Illinois Urbana-Champaign*, USA
 Emily Chen, *University of Southern California*, USA
 Lorenzo Cima, *IIT-CNR*, Italy
 Matteo Cinelli, *Sapienza University of Rome*, Italy
 Daria Dergacheva, *University of Bremen*, Germany
 Roberto Di Pietro, *KAUST*, Saudi Arabia
 Alexandros Efstratiou, *University of Washington*, USA
 Sarah Gilbert, *Cornell University*, USA
 Laurent Hébert-Dufresne, *University of Vermont*, USA
 Natali Helberger, *University of Amsterdam*, Netherlands
 Manoel Horta Ribeiro, *Princeton University*, USA
 Martin Husovec, *London School of Economics*, UK
 Jialun Aaron Jiang, *Meta*, USA
 Paddy Leerssen, *University of Amsterdam*, Netherlands
 Chen Ling, *Indiana University*, USA
 Luca Luceri, *University of Southern California*, USA
 Tanushree Mitra, *University of Washington*, USA
 Anna Monreale, *University of Pisa*, Italy
 Shirin Nilizadeh, *University of Texas Arlington*, USA
 Orestis Papakyriakopoulos, *TU Munich*, Germany
 Antonis Papasavva, *UCL*, UK
 Marian-Andrei Rizoio, *University of Technology Sydney*, Australia
 Giuseppe Russo, *EPFL*, Switzerland
 Martin Saveski, *University of Washington*, USA
 Qinlan Shen, *Oracle Labs*, USA
 Fabrizio Silvestri, *Sapienza University of Rome*, Italy
 Benedetta Tessa, *IIT-CNR*, Italy
 Yannis Theocharis, *TU Munich*, Germany
 Gareth Tyson, *HKUST*, Hong Kong
 Aleksandra Uрман, *University of Zurich*, Switzerland
 Marisa Vasconcelos, *UFMG*, Brazil
 Bertie Vidgen, *The Alan Turing Institute*, UK
 Kai-Cheng Yang, *Northeastern University*, USA

Table 1: Tentative list of TPC members.

matic discussion panel, and a brainstorming and bottom-up discussion session. Each paper presentation will be followed by an interactive Q&A session. Paper session #2 (featuring *resource*, *vision*, and *work-in-progress* papers) will help kickstart the brainstorming and bottom-up discussion, which will focus on providing networking opportunities for participants, stimulating new research ideas and discussions on promising future directions. Some tentative names for the keynote and panel speakers are reported below. The workshop schedule is flexible and will possibly be adjusted based on submissions, speaker availability, and participant needs.

9:00—Welcome.

9:30—Keynote (TBD)—candidates: Robert West, Isabelle Augenstein, Marco Guerini).

10:30—Coffee break.

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

12:30—Lunch break.

13:30—Panel (TBD)—candidates: Adriana Iamnitchi, Dino Pedreschi, Kalina Bontcheva, Stefan Feuerriegel, Claudia Wagner).

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

15:45—Coffee break.

16:00—Brainstorming and bottom-up discussion.

17:15—Closing remarks.

17:30—Conclusion.

Submissions and TPC

We encourage four types of paper submissions: (i) *research* papers, (ii) *resource* papers (e.g., demo, dataset, software),

	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 2: Related workshops.

(iii) *vision* and *position* papers, (iv) *work-in-progress* papers. Each submission will be reviewed by three anonymous reviewers from our experienced and multidisciplinary technical program committee (TPC). Submissions will be reviewed based on their contributions, novelty, rigor, and relatedness to the workshop. Diversity of the organizers and TPC ensures the inclusion of representatives from all relevant disciplines. Table 1 reports the tentative list of TPC members, many of whom have already accepted our invitation.

Other Related Workshops

Table 2 presents a short list of recent related workshops. Our workshop offers a unique and timely expansion to the current landscape of content moderation research by addressing aspects that are overlooked in venues such as CySoc, WOAH, MMCM, Integrity, RegNLP or NLLP. For example, while CySoc centers on cyber social threats, it primarily emphasizes computational approaches for *detecting online harms*, such as disinformation, harassment, and extremist content, with less attention given to the broader content moderation process, which includes community guidelines formulation, development and evaluation of moderation interventions, and moderation impact assessment. In contrast, our workshop adopts a more holistic perspective on content moderation, touching tasks and topics such as policy creation, human factors, platform governance, legal compliance, and the development of quantitative indicators for measuring the effectiveness of interventions. Similarly, WOAH, MMCM, and Integrity workshops are predominantly concerned with *detection mechanisms* and focus on technical solutions grounded exclusively in computer science, as reflected by the composition of their organizing committees and the technical nature of their calls for papers. In turn, workshops such as RegNLP or NLLP cover the NLP analysis of Terms of Service or Community Guidelines as sources of legal obligations, but without a clear connection to content moderation processes and architectures.

Our workshop distinguishes itself by fostering a multidisciplinary approach that integrates computational, social, normative, and ethical viewpoints. This breadth is reflected in our organizing team and TPC, which draws expertise from diverse fields including computer science, social science, and policy studies. The scope of our themes goes beyond detection to critically examine how guidelines are crafted and implemented, the causal effects of moderation interventions, the balance between privacy and harm prevention, and compliance with emerging regulatory frameworks. Such broad and integrated focus makes our workshop a perfect fit for ICWSM—a conference known for its commitment to ad-

vancing research at the intersection of computational and social sciences.

Organizers

The workshop organizers bring together a young, gender-balanced, and multidisciplinary team of scholars with diverse expertise spanning social computing and computational social science, law, human-computer interaction, platform governance, and media policy. Each member is a recognized expert who holds a central position within their respective community, collectively representing a broad spectrum of scholars, perspectives, and interests in content moderation. Below, we provide a short bio for each organizer.

Stefano Cresci is a Researcher at the Institute for Informatics and Telematics of the National Research Council (IIT-CNR) in Pisa, Italy. Stefano's scientific interests broadly lay 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, WebConf, 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 ACM WSDM, IEEE DSAA, and SocInfo.

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, WebConf, 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.

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 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 is an Assistant Professor at Rutgers University. Shagun is an expert in *human-computer interaction* and a pillar in the CSCW and CHI communities. His research focuses on improving content moderation processes on digital platforms from a user-centric, rather than platform-centric, perspective. He critically examines how the design, technical affordances, and policies on large-scale platforms influence the experiences of different stakeholders. His work contributes empirically informed solutions to societal issues such as online harassment and the rise of hateful groups. He also aims to instill fairness and transparency in platform communications with end-users. Shagun co-organized workshops and panels at ACM CSCW.

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 decisions influence content moderation. Her research critically examines transparency, accountability, and the role of public and private governance in shaping online speech. Robyn's extensive experience at the intersection of social science and public policy provides valuable insights into regulatory frameworks, platform accountability, and the societal impact of content moderation strategies. Her work has been featured in top journals and major media outlets, making her an influential voice in ongoing debates about the future of online governance. She recently co-organized workshops at AoIR and the Data & Society Research Institute.

Statement on the use of AI

No generative AI tools were used to generate ideas for the workshop or its activities. All ideas and content in this proposal were developed collaboratively by the organizing team.

References

- Ali, S.; Saeed, M. H.; Aldreabi, E.; Blackburn, J.; De Cristofaro, E.; Zannettou, S.; and Stringhini, G. 2021. Understanding the effect of deplatforming on social networks. In *The 13th ACM Web Science Conference (WebSci)*.
- Bonaldi, H.; Chung, Y.-L.; Abercrombie, G.; and Guerini, M. 2024. NLP for Counterspeech against Hate: A Survey and How-To Guide. In *Findings of the Association for Computational Linguistics (NAACL)*.
- Cima, L.; Trujillo, A.; Avvenuti, M.; and Cresci, S. 2024. The Great Ban: Efficacy and unintended consequences of a massive deplatforming operation on Reddit. In *The 16th ACM Web Science Conference Companion (WebSci)*.
- Costello, T. H.; Pennycook, G.; and Rand, D. G. 2024. Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385(6714).
- Cox, A. L.; Gould, S. J.; Cecchinato, M. E.; Iacovides, I.; and Renfree, I. 2016. Design frictions for mindful interactions: The case for microboundaries. In *The 34th CHI Conference on Human Factors in Computing Systems (CHI)*.
- Cresci, S. 2020. A decade of social bot detection. *Communications of the ACM*, 63(10).
- Cresci, S.; Trujillo, A.; and Fagni, T. 2022. Personalized interventions for online moderation. In *The 33rd ACM Conference on Hypertext and Social Media (HT)*.
- Fortuna, P.; and Nunes, S. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4).
- Gillespie, T. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press.
- Horta Ribeiro, M.; Calais, P.; Santos, Y.; Almeida, V.; and Meira Jr, W. 2018. Characterizing and detecting hateful users on Twitter. In *The 12th International AAAI Conference on Web and Social Media (ICWSM)*.
- Horta Ribeiro, M.; Jhaver, S.; Zannettou, S.; Blackburn, J.; Stringhini, G.; De Cristofaro, E.; and West, R. 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)*.
- Hosseinmardi, H.; Ghasemian, A.; Rivera-Lanas, M.; Horta Ribeiro, M.; West, R.; and Watts, D. J. 2024. Causally estimating the effect of YouTube’s recommender system using counterfactual bots. *Proceedings of the National Academy of Sciences*, 121(8).
- Jhaver, S. 2025. Bans vs. Warning Labels: Examining Bystanders’ Support for Community-wide Moderation Interventions. *Accepted in ACM Trans. Comput.-Hum. Interact.*
- Jhaver, S.; Appling, D. S.; Gilbert, E.; and Bruckman, A. 2019a. “Did You Suspect the Post Would Be Removed?”: Understanding User Reactions to Content Removals on Reddit. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).
- Jhaver, S.; Birman, I.; Gilbert, E.; and Bruckman, A. 2019b. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Trans. Comput.-Hum. Interact.*, 26(5).
- Jhaver, S.; Bruckman, A.; and Gilbert, E. 2019. Does Transparency in Moderation Really Matter?: User Behavior After Content Removal Explanations on Reddit. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).
- Jhaver, S.; Zhang, A. Q.; Chan, Q. Z.; Natarajan, N.; Wang, R.; and Zhang, A. X. 2023. Personalizing Content Moderation on Social Media: User Perspectives on Moderation Choices, Interface Design, and Labor. *Proc. ACM Hum.-Comput. Interact.*
- Kaushal, R.; Van De Kerkhof, J.; Goanta, C.; Spanakis, G.; and Iamnitchi, A. 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)*.
- Kozyreva, A.; Herzog, S. M.; Lewandowsky, S.; Hertwig, R.; Lorenz-Spreen, P.; Leiser, M.; and Reifler, J. 2023. Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences*, 120(7).
- Lazer, D. M.; Baum, M. A.; Benkler, Y.; Berinsky, A. J.; Greenhill, K. M.; Menczer, F.; Metzger, M. J.; Nyhan, B.; Pennycook, G.; Rothschild, D.; et al. 2018. The science of fake news. *Science*, 359(6380).
- Leerssen, P. 2023. An end to shadow banning? Transparency rights in the Digital Services Act between content moderation and curation. *Comput. Law Secur. Rep.*, 48(105790): 105790.
- 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.
- Russo, G.; Verginer, L.; Horta Ribeiro, M.; and Casiraghi, G. 2023. Spillover of antisocial behavior from fringe platforms: The unintended consequences of community banning. In *The 17th International AAAI Conference on Web and Social Media (ICWSM)*.
- Saltz, E.; Barari, S.; Leibowicz, C.; and Wardle, C. 2021. Misinformation interventions are common, divisive, and poorly understood. *Harvard Kennedy School (HKS) Misinformation Review*, 2(5).
- Singhal, M.; Ling, C.; Paudel, P.; Thota, P.; Kumarswamy, N.; Stringhini, G.; and Nilizadeh, S. 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)*.
- Steiger, M.; Bharucha, T. J.; Venkatagiri, S.; Riedl, M. J.; and Lease, M. 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)*.
- Trujillo, A.; and Cresci, S. 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)*.
- Trujillo, A.; and Cresci, S. 2023. One of many: Assessing user-level effects of moderation interventions on r/The.Donald. In *The 15th International ACM Web Science Conference (WebSci)*.
- Trujillo, A.; Fagni, T.; and Cresci, S. 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)*.
- Young, G. K. 2022. How much is too much: The difficulties of social media content moderation. *Information & Communications Technology Law*, 31(1).
- Zannettou, S. 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)*.
- Zhou, X.; and Zafarani, R. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5).