



Why Is Collaboration So Hard?

Roundtable Report

September 2024



The Trust and Safety Foundation (TSF) is a 501(c)(3) non-partisan, charitable organization with the mission to convene stakeholders, share knowledge, and engage in activities that facilitate meaningful advancements in trust and safety.

Why Is Collaboration So Hard?

Roundtable Report

Author

Amanda Menking, Ph.D.

Acknowledgements

TSF is grateful to Nikhil Bhasin for providing the spark—and many of the resources and framing—responsible for this roundtable, and to all of the participants who joined to share their time, expertise, and knowledge. Thank you. TSF is also grateful for a grant from the Future of Online Trust and Safety Fund which allowed Farhana Shahid to attend both TrustCon 2024 and this roundtable.

Recommended Citation

Menking, A. 2024. “Why Is Collaboration So Hard?”. Trust and Safety Foundation.

This work is licensed under [Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International](https://creativecommons.org/licenses/by-nc-nd/4.0/).

Contents

Executive Summary	4
TSF's Next Steps	5
Introduction	6
The Original Problem Space	6
Roundtable purpose and aims	7
The Emergent Problem Space	8
Survey results	8
The Roundtable Discussion	10
Disentangling identities	10
Differing priorities, incentives, and constraints	11
Feelings of frustration and/or powerlessness	11
Unexamined assumptions and expectations	12
Conclusion	13
Caveats	14
References	15

Executive Summary

In July 2024, TSF hosted an in-person roundtable at TrustCon. The initial goal of this event was to bring together a range of stakeholders from academia, the tech industry, and civil society to discuss language support, content moderation, and tooling with the aim of identifying and naming existing gaps, ideating together about possible solutions, and discussing how LLMs and/or AI factor into these present and future considerations. However, the conversations that took place during the roundtable centered more on the results of a survey the participants had taken ahead of time and on the theme of collaboration—and, importantly, why it is so hard.¹

This report provides an overview of the roundtable discussion; describes what the unmet desire survey is, why TSF chose to use it, and how its results helped structure this particular conversation; and shares what was learned so that others seeking to foster multistakeholder collaborations in the T&S ecosystem can build on these insights. Key takeaways include:

- In general, participants were open to both informal and formal collaborations with different stakeholders, with a preference for establishing informal relationships before moving to more formal working relationships.
- Differing priorities, constraints, incentives, and assumptions make collaboration challenging. Oftentimes, participants also find it difficult to disentangle other stakeholders' identities from their organizational affiliations.
- Individuals in the T&S ecosystem often operate within employment arrangements that de-incentivize collaboration with other stakeholder groups, especially those that are external to one's organization/employer.

¹ During a virtual TSF roundtable in May 2024, participants also discussed the challenges of collaboration in the T&S ecosystem; that report can be found [here](#).

TSF's Next Steps

Given the nature of the discussion during this roundtable, TSF has created a mailing list and a Slack channel for participants (and other interested parties) to continue the conversation. TSF is also planning a follow up virtual roundtable in early October 2024.

If you are interested in language support, content moderation, and tooling, are actively engaged in work and/or research in this space, and would like to be a part of future conversations, please [email TSF](#).

Introduction

During TrustCon 2024, TSF hosted an invite-only 120 minute roundtable about language support, content moderation, and tooling. This report provides an overview of the topic, shares key takeaways, and discusses next steps. Although the roundtable discussion was intended to cover language support, content moderation, and tooling (described below as the original problem space), participants shared more about why collaboration in this space—and in the trust and safety (T&S) ecosystem more broadly—is so hard. Therefore, the bulk of this report focuses on the emergent problem space: the challenges of collaboration in T&S.

The Original Problem Space

The vast majority of content on large social media platforms and other websites featuring user-generated content exists in languages other than English, and content moderation is overwhelmingly performed by employees in and from the Global Majority. However, the infrastructure, resources, tooling, and research available to and about these moderation decisions—and the trust and safety workers who do this work—are still available almost exclusively in English. Researchers refer to this disparity as the “resourcedness gap” (Nicholas and Bhatia 2023a), and it is problematic for a number of reasons: from false positives (which can lead to a disproportionate silencing of Majority World voices) to significant gaps in moderation that allow abusive content to remain on a platform to real-world violence (Culliford and Heath 2021; Zelalem and Guest 2021; Nicholas and Bhatia 2023b).

In short, content moderation in a global context performed by a global workforce is a “wicked problem” (Rittel and Webber 1973). Understandably, given the nature of this problem and both the competitive and ever-shifting landscapes of technology companies, what work has been done to address this gap has often been fragmented, siloed, and/or under-funded and under-resourced.

Furthermore, historically, when stakeholders have attempted to address this gap, they have relied primarily on lexicons (which are simply lists of words, or dictionaries). Lexicons, however,

have numerous shortcomings (Gitari et al. 2015; Lorenz 2022; Wang, Day and Wu 2022). As more and more tech companies adopt large language models (LLM) and artificial intelligence (AI) to inform and perform content moderation tasks (Markov et al. 2023), the current challenges, gaps, and shortcomings already inherent in lexicons are likely to be compounded (Singh 2019). However, LLMs and AI, if built with this problem space in mind, may also offer the possibility of more responsive, dynamic tooling—particularly if there are affordances that allow language communities to build their own solutions (Rani et al. 2020; Jhaver et al. 2022).

Roundtable purpose and aims

The purpose of this roundtable was to bring together a range of relevant stakeholders to talk about the issues surrounding content moderation when it comes to language support and tooling, with an eye to identifying and naming existing gaps, ideating together about possible solutions, and discussing how LLMs and/or AI factor into these present and future considerations. Importantly, **this roundtable was also intended to provide an opportunity for a range of stakeholders to understand how others are situated in terms of incentives and constraints with the aim of beginning to build collaborative relationships** of respect and trust.

The Emergent Problem Space

To set the scene, it is important to note this roundtable was held early in the afternoon of the last day of TrustCon, the annual conference for trust and safety professionals. TrustCon 2024 saw more than 1,200 attendees across seven tracks. TSF invited 36 participants to attend this roundtable: 20 accepted and 17 participated. Of the 17 who participated, all but three attended the conference.²

The roundtable began with introductions, including names, pronouns, affiliations and/or relevant work, in addition to a participant-suggested icebreaker question about one's favorite tree. The icebreaker question, in particular, engendered reflections and conversations about geography, memory, and sense of place.

TSF's Research & Program Director then shared a prepared deck presenting the results from a survey completed by roundtable participants ahead of the event. Discussion of these results and relevant questions and comments occupied all but 30 minutes of the remaining time. (That is, the roundtable only began to address the problem space during the last half hour of the roundtable.)

Survey results

Ahead of the roundtable, 21 respondents answered the unmet desire survey, a tool designed by Adam Seth Levine to foster new collaborative relationships between different stakeholders (Levine 2022). Of the 21 total responses, four were provided by people affiliated with academia, 10 by people affiliated with the tech industry (and, more specifically, with trust and safety work), five by people affiliated with nonprofit organizations, and two by people who selected "Other" as a stakeholder group. Of the 21 responses, 17 were complete (all questions were answered).

² Roundtable participants did not represent their organizations/companies.

In addition to identifying their affiliation, respondents were asked the following questions:

- Are you looking for informal collaborations (oriented toward knowledge exchange) or formal collaborations (oriented toward projects with shared ownership, decision-making authority, and accountability) with academics, civil society members, and/or industry professionals?
- What hesitations (perhaps due to prior experiences, lack of explicit permission, stereotypes, etc.) do you have about interacting with academics, civil society members, and/or industry professionals?
- What hesitations do you think they might have about interacting with you?
- Why do you think these collaborations don't already exist? (If you already have these collaborations, please tell us more about them.)
- Please share anything else you'd like us to know.

In general, **all respondents were open to both informal and formal collaborations**, though respondents affiliated with the tech industry leaned more towards informal collaborations at first; as one respondent noted, “informal collaboration primarily, however, if appropriate I'd be open to maturing partnerships towards a more formal collaboration.”

Eleven of the 17 respondents didn't have any hesitations about interacting with collaborators affiliated with other stakeholder groups. Of those who did have hesitations, those affiliated with the tech industry were more concerned with **accountability** (for informal collaborations), **legal constraints** (for formal collaborations), **lack of trust**, and **the challenge of discussing constraints** in industry. Respondents affiliated with nonprofit organizations had hesitations about **attribution** and **power dynamics** (especially when collaborating with academics), and both respondents affiliated with academia and nonprofit organizations expressed a concern that **priorities may not be aligned**, specifically when it comes to the Global South and non-English languages: “I feel Global South is deprioritized both within industry and civil societies based in the US.”

In general, respondents from all stakeholder groups indicated they think others may be hesitant to interact with them due to either a **lack of awareness** (“they don't know me”) or a **lack of**

trust. As one respondent affiliated with a nonprofit organization noted: “Seeing civil society as a lofty or not making relevant arguments to the constraints industry actors face internally. I think that's valid. The challenge is we don't have a lot of visibility into the constraints or levers we do have to advocate.”

When asked why these collaborations do not already exist, a few respondents replied that they do and shared specific examples, including [The Trust & Safety Teaching Consortium](#) and [Wikimedia research projects](#). However, overwhelmingly, respondents indicated that **different constraints, incentives, interests, and timelines in addition to a lack of resources and working in silos** all contributed to *preventing* these kinds of collaborations. As one respondent aptly noted, “Collaborations between academics, civil society, and trust and safety industry experts are not as widespread as they could be due to several factors: Different priorities and incentives, communication and language barriers, conflicting interests as industry may be hesitant to share sensitive data or proprietary information, but specially resource constraints.”

The Roundtable Discussion

The majority of time during the roundtable discussion was devoted to discussing the survey results outlined above; this roundtable became much more about the question, “Why is collaboration so hard?” than about the proposed topic—language support, content moderation, and tooling. **So, why is collaboration in this space and across these stakeholder groups so hard?**

Disentangling identities

Roundtable participants talked about **how difficult it is to disentangle individual and organizational identities, incentives, and responsibilities**—both in their own work and when considering potential collaborators. For example, one person remarked, “I can trust someone as a person, but not in their role within an organization.” This theme relates to themes below and underscores the importance of clearly articulating when someone is engaging in a collaboration as an individual or on behalf of an organization or employer.

Differing priorities, incentives, and constraints

As evidenced in the survey responses and in the roundtable discussion, it is apparent that **different priorities, incentives, and constraints contribute to the challenge of collaboration**. But what does this mean exactly?

- Priorities, in this discussion, seemed to refer to **organizational or institutional priorities** (for example, a platform seeking to make a profit, or a civil society organization seeking to make change).
- Different **systems of incentives are inextricably linked to organizational or institutional priorities**. For example, intellectual contribution is the currency of the academy; therefore, many academics (specifically those who have not yet been granted tenure in the U.S. system) are incentivized to spend their time working on formal collaborations that result in peer-reviewed publications and/or seeking access to data so that they can publish relevant studies.
- Constraints may refer to **legal constraints** (such as non-disclosure agreements, terms of data and/or privacy protection), **financial constraints** (one survey respondent noted, “One small problem is people fighting for the same point of money.”), and/or **temporal constraints** (the same respondent continued, “The main reason though is just the effort needed in getting such collaborations going. It distracts away from the person's main organization especially if a financial runway is absent.”)

Feelings of frustration and/or powerlessness

In listening to the discussion and reviewing the notes taken by Maia Levy Daniel and Farhana Shahid, a theme emerged that can best be described as “feelings of frustration and/or powerlessness.”

On the one hand, **participants who have been unable to gain access to the information and/or data they needed** (for example, participants affiliated with nonprofit organizations reported not knowing which “levers” to pull when advocating on behalf of a person or

population to a tech company) expressed these feelings. On the other hand, **participants who have been unable to share information** (for example, trust and safety professionals who were treated as representatives of their employers rather than as professionals with specific expertise across past and current employers), or **whose work has been directly, negatively impacted by the actions of those who pursued other means** of exerting pressure because they themselves felt powerless also expressed these feelings.

One particularly powerful example provided by a participant that resulted in discussion: Public “naming and shaming” initiatives undertaken by stakeholders outside of tech companies (who feel frustrated and/or powerless because groups in the Majority World on behalf of whom they advocate may not have access to contacts, representatives, and/or a partnership program within a given company) have often been counterproductive in that they inject chaos and result in panic and marginal change—leaving trust and safety professionals inside of said tech companies feeling frustrated and/or powerless.

Unexamined assumptions and expectations

During the discussion about public “naming and shaming,” a trust and safety professional remarked there’s a **“profound misunderstanding”** about the profession and the role(s) of trust and safety teams within the larger organizational structures of tech companies, saying it is simply a “decision-widget machine” in many respects. This observation resonated with other participants affiliated with the tech industry, some of whom commented there is a “tendency to techno-solutionism” and an expectation “engineers [can] solve ethics questions.”

Relatedly, a participant affiliated with a university shared they often have more disagreements with other academics than with external stakeholders, pointing out that academics are not as monolithic as people outside of the academy may assume.

Conclusion

Woven throughout the survey responses and the roundtable discussions is a clear thread: **individuals in the T&S ecosystem often operate within employment arrangements that de-incentivize collaboration with other stakeholder groups, especially those that are external to one's organization/employer.** This may not always be explicit (such as in the form of a nondisclosure agreement); it can be implicit too (such as in the form of the time constraints or unexamined assumptions and expectations about other stakeholder groups).

This thread resonates with Levine's (2022) observation in working with local policymakers and academic researchers on the topic of informal collaboration (defined as, "a collaborative exchange in which people with diverse forms of knowledge, expertise, and lived experience share what they know with the goal of developing an expanded understanding of a problem—yet still remain autonomous decisionmakers"):

As a result, I see an important yet currently missed opportunity for researchers to become more involved in policy. But doing so will require both institutional support and introspection from researchers themselves. First, institutions and funders must assist and incentivize researchers to initiate and maintain these valuable informal collaborations with local policymakers. Second, researchers must learn more about these policymakers and their needs, and how their expertise can meet those needs.

Moving forward, TSF will continue to prioritize providing opportunities for stakeholders in the T&S ecosystem to come together to better understand one another and to begin to lay the groundwork for both informal and formal collaborations. In particular, TSF will specifically focus on building trust among participants—not only from this roundtable discussion but also from other collaborative meetings and conversations—to ensure a safe space for everyone to share their expertise, knowledge, and lessons learned in the field. This trust will be imperative to raise awareness of the importance of collaboration in the T&S space, as well as to start to address

the institutional barriers mentioned above and to create the adequate incentives for different stakeholders to work together.

Caveats

TSF's goal in sharing this report is not to establish a baseline or descriptive analysis of any kind (in fact, TSF's goal was *not* to conduct research), but rather to contribute to and support future collaborations in this space. TSF recognizes there was a self-selection bias at play: survey respondents and roundtable participants wanted to collaborate with other stakeholders. As a research "product," the unmet desire survey results and the roundtable discussions described above are exploratory at best.³

³ Additionally, qualitative data is not meant to be generalizable, but rather trustworthy—credible, transferable, dependable, and confirmable.

References

- Culliford, E., and Heath, B. 2021, October 26. "Language Gaps in Facebook's Content Moderation System Allowed Abusive Posts on Platform: Report." The Wire.
<https://thewire.in/tech/facebook-content-moderation-language-gap-abusive-posts>
- Jhaver, S., Chen, Q. Z., Knauss, D., and Zhang, A. X. 2022, April. "Designing Word Filter Tools for Creator-Led Comment Moderation." In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 1-21.
- Gitari, N. D., Zuping, Z., Damien, H., and Long, J. 2015. "A Lexicon-Based Approach for Hate Speech Detection." International Journal of Multimedia and Ubiquitous Engineering, 10(4), 215-230.
- Levine, A. 2022. "Unmet Desire." Issues in Science and Technology 38(3), 27–31.
- Lorenz, T. 2022, April 8. "Internet 'Algospeak' is Changing Our Language in Real Time, from 'Nip nops' to 'Le dollar bean'." The Washington Post.
<https://www.washingtonpost.com/technology/2022/04/08/algospeak-tiktok-le-dollar-bean/>
- Markov, T., Zhang, C., Agarwal, S., Eloundou Nekoul, F., Lee, T., Adler, S., Jiang, A., and Weng, L. 2023. "A Holistic Approach to Undesired Content Detection in the Real World." Proceedings of the AAAI Conference on Artificial Intelligence, 37(12), 15009-15018.
- Nicholas, G., and Bhatia, A. 2023a. "Toward Better Automated Content Moderation in Low-Resource Languages." Journal of Online Trust and Safety, 2(1).
<https://doi.org/10.54501/jots.v2i1.150>
- Nicholas, G., and Bhatia, A. 2023b. "Lost in Translation: Large Language Models in Non-English Content Analysis." Center for Democracy & Technology.
<https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/>
- Rani, P., Suryawanshi, S., Goswami, K., Chakravarthi, B. R., Fransen, T., and McCrae, J. P. 2020, May. "A Comparative Study of Different State-of-the-Art Hate Speech Detection Methods in Hindi-English Code-Mixed Data." In Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, 42-48.

Rittel, H.W. and Webber, M.M. 1973. "Dilemmas in a General Theory of Planning." Policy Sciences, 4(2), 155-169.

Singh, S. 2019. "Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content." New America, 22, 1-42.

Wang, C. C., Day, M. Y., and Wu, C. L. 2022. "Political Hate Speech Detection and Lexicon Building: A Study in Taiwan." IEEE Access, 10, 44337-44346.

Zelalem, Z. and Guest, P. 2021, November 13. "Why Facebook Keeps Failing in Ethiopia." Rest of World. <https://restofworld.org/2021/why-facebook-keeps-failing-in-ethiopia/>