Behavioural Scripts and the Effect of User Moderation on Social Media

Julian Dehne[‡]
Valentin Gold[§]

June 9, 2023

Abstract

This paper analyzes user moderation in social media. Focusing on online civic interventions, it analyzes user reactions to these interventions, incorporating behavioral scripts and users' previous behavior. The study draws from deliberation theory, script theory, and the sociology of knowledge, examining the assumption that user moderation leads to productive or counterproductive patterns of reactions in rational discourse. The research aims to strengthen deliberation in social media amidst challenges such as automated trolls, misinformation, and competing interests.

1 Introduction

With social media being in a danger of loosing its deliberative impact with automated trolls, strategic misinformation and marketing interests competing with scientific evidence or ethic-based values, it is paramount that proactive methods of strengthening deliberation in social media are found.

For many years, content moderation on social media was restricted to deletion; users were only able to report content for manual review thus creating a large workload for human moderators. However, recent developments in the field of artificial intelligence open up possibilities to go beyond this traditional approach: AI moderators can be trained to act on their own by first determining when a conversation starts to deteriorate and then to decide on the type of intervention needed. In this paper, we are laying the foundation for the latter direction of research: we conduct a large-scale analysis of Tweets and Reddit posts with the question of how users react towards attempts of moderation. In particular, we focus on one type of intervention: online civic interventions - a specific kind of metacommunication adopted by ordinary users to moderate their peers (Porten-Cheé et al., 2020). Given that a user has intervened and pointed out

[‡]Georg-August Universität Göttingen
[§]Georg-August Universität Göttingen

a misconduct of a conversational norm, for instance by telling other users to be more friendly, we analyze the reactions towards these interventions. Our approach is based on the idea of behavioral scripts that guide users in their online communication (Kluck and Krämer, 2022). Consequently, we also include users' previous behavior in order to explain their reactions. Based on our results, we suggest a set of (personalized) interventions that are most likely to succeed in improving inappropriate online behavior.

As a theoretical framework we are drawing from deliberation theory (mainly Habermas), script theory and the sociology of knowledge by Mannheim and its application within the documentary method. Using these theoretical frameworks a concept of deliberative moderation is conceived.

Different strategies of deliberative moderation are investigated. An emphasize was placed on creating a corpus that includes strategies of user moderation where coders had a high agreement. Walking the line between a clear concept of moderation and one that actually happens "in the wild" a definition of partial moderation was derived.

The assumption being investigated in this paper is that user moderation – although not intended as a deliberative force– leads to patterns of reactions that can be categorized as productive or not. The productivity is analyzed with regard to the rationality of the discussion that ensues after the moderating intervention took place.

2 Related Work

There are directions of research that play a role for this study: first, the question when users write moderating posts and second what the reactions are. The first question is relevant for the sampling process. The clearer user moderation can be characterized the easier it is to automatically mine moderating posts using computational methods. The second question is relevant in order to compare the findings of this paper with other ways to model user reactions and their impact on deliberative quality.

Moderation can be classified as a specific type of metacommunication. Previous research shows that the occurrence of metacommunication among individual users is influenced by various factors related to online discussions. One significant factor is the perception of threats to favorable conditions for political online discourse.

According to Porten-Cheé et al. (2020), when such threats are perceived, it triggers a response in other users. Furthermore, the definition of what constitutes a norm violation is determined solely by the participants actively engaged in the discussion. Porten-Cheé et al. (2020) highlight that ordinary users have the authority to identify and label norm violations within the discourse. According to Bormann et al. (2021), among the different types of norm violations, those

related to context norm and relation norm tend to be considered more severe compared to others. They assert that violations of political context norms, relation norms, information norms, process norms, and modality norms are perceived as particularly significant.

It is important to note that the consequences of uncivil user comments vary depending on the social context. Kluck and Krämer (2022) emphasize that the impact of uncivil comments is strongly influenced by the specific social dynamics at play. In addition, the severity of an attack or violation matters to the average participant in the discussion. Pradel et al. (2022) suggest that the perceived frequency and severity of norm violations directly affect the likelihood of individual users engaging in metacommunication.

In conclusion the previous research shows that the more frequent and severe perceived norm violations are, the more likely individual users will engage in metacommunication.

Overall, there are three rule-related social processes in online communities: rule-breaking, rule-making, and rule-enforcement (Sternberg, 2012, 155-169). Moderators in Twitch communities, as identified by Wohn (2019), can assume different roles. These roles include Helping hands, Justice Enforcers, Surveillance Units, and Conversationalists. Cai and Wohn (2019) identify five approaches that moderators may take to dealing with problematic behaviors: Educating, Sympathizing, Shaming, Humor, and Blocking.

Seering also refers to Kraut et al. (2012) who outline five key difficulties that community leaders encounter: (1) Encouraging Contribution; (2) Encouraging Commitment; (3) Regulating Behavior; (4) Dealing with Newcomers; and (5) Starting New Communities. However, (Seering, 2020, 18) comes to the conclusion that the effectiveness in moderation cannot be defined universally.

Molina and Jennings (2018a) hypothesize that comments condemning the incivility of other commenters would result in more civil comments and a decrease in uncivil comments. However, their findings did not yield significant results to support this hypotheses. Although they did not observe a significant impact of metacommunication on the overall civility level, which would have indicated the participants' ability to independently change the tone of a conversation without relying on moderators, they found evidence of metacommunication being practiced. This suggests that commenters positively responded to individuals attempting to intervene and denounce incivility.

Other researchers conclude that metacommunication (e.g., talking about the tone of discussion) engendered more metacommunication and is usually caused by an uncivil discourse (Han et al., 2018, 1) However, metacommunication does not significantly decreases incivility, but rather increases metacommunication, according to Molina and Jennings (2018b, 56).

Porten-Cheé et al. introduce a new concept of political participation: Online Civic Intervention (OCI). They "define OCI as the action taken by ordinary

users to restore favorable conditions for political online discussions when threats to these conditions are perceived" (Porten-Cheé et al., 2020, 519) . They point out that it is important to distinguish OCI from mere expression of opinions. Rather, OCI represents a form of metacommunication aimed at promoting a more civil and rational public debate, with the goal of ensuring an inclusive online discourse. Furthermore Porten-Cheé et al. (2020) argue that OCI serves to restore an atmosphere of mutual respect and reliable information, creating a suitable environment for discussing political matters while allowing for dissenting viewpoints and freedom of speech.

Importantly, OCI embodies a highly democratic expression since the determination of what constitutes relevant and civil discourse is made exclusively by the ordinary users actively engaged in the discussion, rather than being dictated by a select group of professionals who set the terms of the platform..

According to Watson et al. (2019), younger, wealthier, white males are more likely to report abusive comments. Additionally, trust in the news media and authoritarian personality traits were also found to be significant predictors of bystander intervention. Also, research has found that users who are more often exposed to uncivil comments are more likely to make uncivil comments themselves (Han et al., 2018, 23).

Furthermore, according to Kluck and Krämer (2022), people were more likely to attribute aggressive motives to senders of incivility when they opposed their opinion. In a study conducted by Pradel et al. (2022), researchers found that among the different scenarios presented, violent threats resulted in the highest level of support for content moderation measures. These measures included actions that might be considered censorship in certain contexts, such as removing content or suspending user accounts. However the research also showed that there is generally a low level of demand for addressing offensive behavior and the average participant's level of support depends on the severity of the attack. Also, Molina and Jennings (2018b, 57-58) found that when users "were made aware of the uncivil tone in a conversation, they were more likely to try to intervene in the situation and point out the incivility they witnessed".

Data collected by Kluck and Krämer (2022) shows that preceding comments have a limited influence on the perception of a comment. Users were more inclined to attribute aggressive motives to a commentator when they had opposing opinions. The attribution of aggressive motives "increased individuals' anger, anxiety, hostile cognition's, but also enthusiasm" (Kluck and Krämer, 2022, 1). The General Aggression Model (GAM), proposed by Anderson and Bushman in 2002, suggests that individuals develop specific knowledge structures related to aggression based on their experiences. There are three specific knowledge structures that particularly influence the interpretation of aggression in social situations: perceptual schemata (such as recognizing an uncivil comment), person schemata (beliefs about the comment sender), and behavioral scripts (information on how individuals should respond to a discussion comment). Drawing on this, Kluck and Krämer (2022, 2) conclude that, when individuals

perceive a comment from other users as uncivil, their response is influenced by behavioral patterns they have learned from previous experiences. Nevertheless, categorizing a comment as uncivil can be a complex process. The focus theory of normative conduct (Cialdini, 2003) proposes that there are two different types of norms: descriptive norms, that refer to how individuals perceive typical behavior and influence their actions by presenting evidence of what is likely to be effective and adaptive, and injunctive norms, that refer to assumptions about behaviors that are commonly accepted or disapproved by others (Kluck and Krämer, 2022, 6).

Taking all these findings into consideration, it becomes apparent that a general theory of deliberative moderation is missing. Moderating behaviour is mainly researched as part of a bundle of reactions to uncivil behaviour rather than as a proactive force. It is still uncertain whether moderation shares the same preconditions and effects as general metacommunication, or if individual characteristics can be observed.

3 Theory and Hypotheses

In the next section moderation and deliberation are defined and tied together. Drawing on political philosophy and psychological research on moderation and mediation, a framework is developed that situates moderation between the deliberative function and its practice in ongoing social scripts.

Moderation is then operationalized as a set of strategies and expressions of these strategies so that it can be found and labeled in social media. Finally, a discourse-analytical framework is imported from qualitative social sciences in order to categorize the effect moderation has in terms of discourse component patterns. These can be interpreted as social scripts or simply as patterns of reaction to moderation.

3.1 Deliberation and Moderation

Moderation and deliberation are two concepts that can be closely linked. In the context of this study social media is viewed from the perspective of how online communication may foster the exchange of views and arguments, improve or worsen the social bonds that hold together democratic societies.

Moderation, as an essential component of deliberation, plays a vital role in fostering constructive and inclusive discussions. Deliberation, at its core, is the process of exchanging ideas, perspectives, and opinions to reach informed decisions. According to Stromer-Galley (2007, 3)

> *deliberation is defined [...] as a process whereby groups of people, often ordinary citizens, engage in reasoned opinion expression on a social or political issue in an attempt to identify solutions to a common problem and to evaluate those solutions.*

It is through moderation that we ensure this exchange occurs in a respectful, fair, and productive manner. Moderation acts as a guiding force, steering deliberations towards their intended purpose – to explore diverse viewpoints, challenge assumptions, and ultimately arrive at well-considered outcomes. By setting clear guidelines and facilitating a respectful atmosphere, moderators enable participants to engage in meaningful dialogue, fostering an environment where ideas can flourish.

One of the primary functions of moderation within deliberation is to ensure equitable participation. Moderators create a space where all voices are heard, irrespective of their background, identity, or beliefs. By actively encouraging marginalized or underrepresented perspectives to be shared and giving them equal weight, moderation helps prevent dominant narratives from overshadowing valuable contributions.

Moderators also play a crucial role in managing the tone and civility of discussions. They help maintain a respectful discourse by discouraging personal attacks, derogatory language, or any form of harassment. By promoting empathy and understanding, moderators cultivate an environment that fosters active listening and empathy, allowing participants to engage with differing viewpoints constructively.

Furthermore, moderation enhances the quality of deliberation by ensuring adherence to evidence-based reasoning and critical thinking. Moderators can help identify and address misinformation, fallacious arguments, or logical inconsistencies, thereby raising the overall standard of the discourse. By promoting thoughtful analysis and fact-based dialogue, moderation contributes to informed decision-making.

Ultimately, moderation and deliberation go hand in hand, as moderation supports the very essence of deliberative processes. By promoting equitable participation, maintaining a respectful tone, upholding evidentiary standards, and encouraging active engagement, moderation strengthens the deliberative process, enabling diverse voices to shape meaningful outcomes. According to Chambers (2003, 318) there is the belief that deliberation will improve the healthiness of public discussions.

On the one hand, these ideas reflect a normative perspective where moderation plays the defined role of improving the rationality of the conversations or leads to consensus building or meaningful exchange of ideas. On the other hand, following a pragmatist perspective, moderation is part of ongoing (cultural) scripts that are semi-stable and from which humans derive their repertoire of action.

3.2 Normative and Pragmatist Concepts of Moderation in Social Media

Earlier idealistic views on the internet focused on the relative freedom of speech in social media and the chance of fostering international publics with the means of the new technologies. With the rise of hate speech and artificial destructive agents these notions have to be put into perspective. Less attention has been given to the assumptions of rationality, productive talk, thematic focus and consensus building. These are aspects of a public sphere that need to exist in one way or another to fulfil its task as a channel for social advancement.

We argue that some discussions in social media constitute a discursive type of public space (*Öffentlichkeit*) – but this is not the case for all communication on these platforms. The prerequisite for our research in this paper is that there are some aspects of social media that can be interpreted as both rational and generally accessible.[1] Habermas argues that these two requirements for a deliberative discourse are in conflict with each other:

> *The moment of publicity, which guarantees rationality/reasonableness, is to be preserved at the cost of its other moment of generality, which guarantees general accessibility [translated by the authors].* (Habermas, 1982, 346)

It is noteworthy that the German word "Vernünftigkeit" translates to both rationality and reasonableness. Whereas rationality might be too high an expectation, transforming communication in social media to a level of reasonableness should be within reach.

In his Theory of Communicative Action, Habermas (1981a) introduces a more general framework from which to derive the concept of a discourse which does not require the public sphere as a liberal concept but which defines communicative action based on rationality:

> *A conversation is a discourse if and only if a rational motivated consensus could be achieved assuming that the arguments can be put forward as often as necessary and for as long as necessary* (adapted and translated by the authors) (Habermas, 1981a, 71)

In contrast to the concept of the public sphere, this concept of communication fits perfectly to the conversations held in social media as they do not impose any restrictions as to how often arguments are presented.

Contrary to live debates, conversations held in social media are not dependent on physical conditions of humans becoming fatigued and can stretch time indefinitely – assuming the platform stays online and the *attention* does not

[1] We mainly build upon the definition of rationality as laid out by Habermas. However, we limit ourselves to exactly this contribution by Habermas and we do not extend our definition to a liberal theory of deliberation. To make this very clear, social scripts, cultural scripts and patterns are included in the theoretical groundings, attempting to merge the Habermasian idea of rationality with a pragmatic approach to deliberation.

waver. The first part of the definition introduces the consensus as a criteria for a discourse differentiating a conversation from a discourse *by the result*.[2]

The following quote links consensus building to the single participants, displaying the efforts of Habermas to integrate systems theory with symbolic interactionism and Durkheim's theory of solidarity (Habermas, 1981b, 19):

> A discoursive consensus depends at the same time on the yes and the no of each single participants, and it depends on surpassing egocentrism. (translated by the authors) (Habermas, 2009, 19)

This state of consensus is hard to evaluate empirically and there is a clear bias to focus on the outcomes as they translate to easy-to-measure variables. Moreover, it is very unlikely that discussions in social media resemble this ideal of a discourse which – in its nature – is a thought experiment and not an empirical heuristic.

Habermas postulates that there is a communicative rationality that can be derived from the need to coordinate actions based on existing norms or the human need of expressing oneself. However, social media like Twitter or Reddit seem to be subsystems that do not require a consensus as there is no immediate action tied to it that needs to be coordinated.

In summary, from the Habermasian theory we use the concept of communicative rationality. It is important for deliberation and the public sphere to assert both reasonableness and accessibility without prioritizing one over the over. Moreover, we assume that discussions in social media can be (symbolic of) discourses as they are open for anyone to access and, in principle, arguments can be exchanged indefinitely.

One of the arguments against measuring deliberative quality is the nature of internet discussions, and equally important the nature of human talk in general which most often do not comply to the high normative standards of rationality. Consensus may just not be possible in an online setting where identity is defined differently and the conditions that encourage compromise are lacking (Papacharissi, 2004, 269). Moreover, researchers focus on social scripts, "'the violation of which leads to flaming"' (Papacharissi, 2004, 269). However, the emotional guise should not be mistaken for the absence of norms. As Papacharissi (2004, 281) point out, "[i]ncivility [...] is fundamentally linked to attitudes and beliefs, and as such could have graver repercussions [than impoliteness]".

When evaluating a discussion, it is important to differentiate between its quality based on ethical or democratic ideals and its emotional and complex appearance.

[2]From this it follows that offering arguments alone is a necessary but not a sufficient condition for deliberative quality but for the discussion becoming a discourse. In the following the words discussion and discourse are used interchangeably for this reason as we have defined a discourse as a discussion based on rational argument and with reference to the Habermasian formula.

A discussion that is free from power dynamics and conducted with respect, following the principles of non-violent communication (Rosenberg, 2016), can help prevent emotions from escalating, rudeness, and even hatred. Although human behavior on social media often prioritizes saving face over rationality, we should still consider rationality as the benchmark for assessing the quality of these discussions, assuming that rationality allows for the presentation of factual information.

However, there are many suppliers for believe systems that shape the perception of participants in social media that are outside of the scientific world and do not adhere to its standards.

> *A diverse combination of actors including trolls, bots, fake-news websites, conspiracy theorists, politicians, highly partisan media outlets, the mainstream media, and foreign governments are all playing overlapping–and sometimes competing–roles in producing and amplifying disinformation in the modern media ecosystem.* (Tucker et al., 2018, 32)

Finally, the question of measuring deliberation empirically has been stomped by Bächtinger and Parkinson:

> As an empirical theory, deliberative theory has been widely criticized for making assumptions that seem to fly in the face of what scholars already know about human behavior. (Bächtiger and Parkinson, 2019, 50). [. . .]

> In the light of our new problem-based approach to conceptualizing deliberation and deliberativeness, such goals are misdirected. In short: we declare that the time of searching for a grand, unified index of deliberative quality are over. (Bächtiger and Parkinson, 2019, 138).

As an alternative, Parkinson et al. (unpublished manuscript) are investigating cultural scripts to empirically access deliberative processes without using the normative judgement based on critical theory. Cultural script theory stands in the tradition of pragmatist philosophy and assumes that a reply in social media is based on a "toolkit" of social scripts and can be researched by analyzing certain patterns of script sequences.

For instance, a moderating statement may be part of the script "knowledge exchange" that consists of the steps (a) asking for elaboration, (b) answering and elaborating, and (c) differentiating the answer, with (a) being the moderating intervention in the discussion.

The position in this paper incorporates both perspectives. On the one hand rationality is deemed important for deliberative quality: it can be used to derive social norms and it is useful for consensus building. On the other hand, using the concept of Habermas' power-free discourse as an empirical measurement

is rejected and replaced by the idea that it is closer to the messiness of real communication to focus on the patterns of communication witholding a final judgement.

Assuming that moderation is useful for deliberation, this approach will focus on the effect of different moderation patterns rather than its effectiveness in terms of deliberative quality overall. However, by using a discourse-analytical framework there still remains a focus on rationality (arguments, contradictions and conclusions) when analyzing the patterns. This means, that some of the empirical data will need to be disregarded as they do not contain any rational discussion and cannot be coerced into a closest-fit category.

3.3 Defining User Moderation in Social Media

According to Grimmelmann (2015, 47), moderation in social media can be defined as "the governance mechanism that structures participation in a community to facilitate cooperation and prevent abuse". The ultimate goal of moderation is to foster cooperation among community members, allowing them to build common ground and reach consensus on various issues. This is achieved through the implementation of governance mechanisms, which serve as the means to accomplish this goal.

In the next sections moderation will be operationalized using these steps: first, different functions of moderation are discussed based on the literature and the theoretical embedding within deliberation. Limiting the functions of interest to the process function different social contexts are discussed with regard to what moderation looks like. The different social contexts of moderation are then modelled according to the focus and the dispositions of the participants in the conversation. Finally, the deliberative functions of moderation, the process function and the social contexts are integrated abductively into different strategies of moderation. These strategies are represented by phrases that are used to filter the data available in social media in order to find candidates for moderation.

3.3.1 The Function of Moderation

As proposed by Edwards (2002, 8), the function of moderation in social media can be understood through three key aspects. Firstly, the strategic function of moderation involves establishing the boundaries of the discussion and embedding it within the political and organizational context of the community. This function helps shape the direction and scope of the discussion, ensuring it aligns with the overarching goals and values of the community.

Secondly, moderation serves a conditioning function by translating the strategic outline into various conditions and provisions for the discussion. This includes setting rules, guidelines, and policies that govern user behavior and content. By implementing such conditions, moderation creates a structured environment

that promotes productive and respectful interactions among community members.

Lastly, moderation fulfills a process function, which encompasses all tasks related to the discussion process itself as a collective and purposeful activity. This includes tasks such as facilitating dialogue, mediating conflicts, encouraging active participation, and promoting the overall quality of the discussion. By performing these process-related functions, moderation ensures that the discussion remains constructive, inclusive, and conducive to achieving the goals of cooperation and consensus-building.

Whereas participation in a community is more abstract, the three functions by Edwards assume the discussion as the mode of implementing participation. This is not the only possibility but it is the one we are concerned with. The strategic function and the conditioning function describe aspects that take place before a discussion or in the case of the strategic function sometimes within a discussion but with a perspective that transcends the agenda such as enforcing disciplinary boundaries in a university course. For this reason, we are only concerned with the process function as it is the only one that applies to social media.

In terms of characterizing moderators, Friess (2021) differentiates between professional moderators, user moderators, ordinary users and collective civic action. Here we are only concerned with ordinary users that take on the role of a moderator in addition to their role as participants, interested parties etc. This means, that moderation will not be disjoint from other aspects of the message these users are sending. Also in the case of user moderators or professional moderators the norms they are enforcing are pre-defined. For instance, in a Subreddit the rules for this Subreddit may be written down explicitly. The user moderators would then take on the task to decide whether a situation is at odds with these rules and intervene by deleting the post or writing a moderating statement, warning the user that s/he has breached the rules.

For ordinary users, the motivation to write is not to enforce social norms. Social norms may come out involuntarily or as a rhetorical device. For this reason – and based on the definition by Edwards (2002) – partial moderation is defined as the sentence within a post or tweet that appeals to a social norm that tries to structure participation in order to foster cooperation within the discussion process itself as a collective, purposeful activity that fits the social context of the situation. In order to label a post as moderating partial moderation is considered a sufficient requirement[3].

3.3.2 The Social Context of Moderation

The social context of the situation can be differentiated according to the type of community and the type of cooperation that is possible and expected within

[3]In a preliminary study, full moderation (every sentence of the post was in some way moderating) was only found at a ratio of approximately 1/10000 (is moderating/moderation candidate).

that community. For example, in an academic discussion the cooperative mode would be a rational argument with a focus on consensus, truth seeking and a strong focus on the issue at hand. With a less strong focus on arguing, in a diplomatic environment like the United Nations Assembly or a trade dispute, the cooperative mode would be bargaining and trying to find an optimal compromise. Finally, in a couples' therapy the cooperative mode would be to discuss the relationship and bring in all the emotions involved.

One of the main difficulties with this concept is with regard to the social media setting – it is not an easy task to apply this concept to social media. In general, there are some relevant questions like what is the (sub-)community like? We even don't know what we should expect with regard to the level of cooperation within the community. Also, an open question is how closely the community is linked and how clear-cut are the fringes of the community?

Because these questions cannot be answered (where we are at), all kinds of attempts that look like invoking a social norm can be labeled as moderating as long as they are meant to structure the participation.

Based on Black (2013) some types of communication conflicts can be viewed as storytelling with different dispositions (frames). On the one hand there is the difference between unitary and adversarial concepts of democracy, on the other hand conflicts may be more or less about relationships or issues. In Fig 1, using these combinations of dispositions as dimensions, we derive at a context model for moderation.

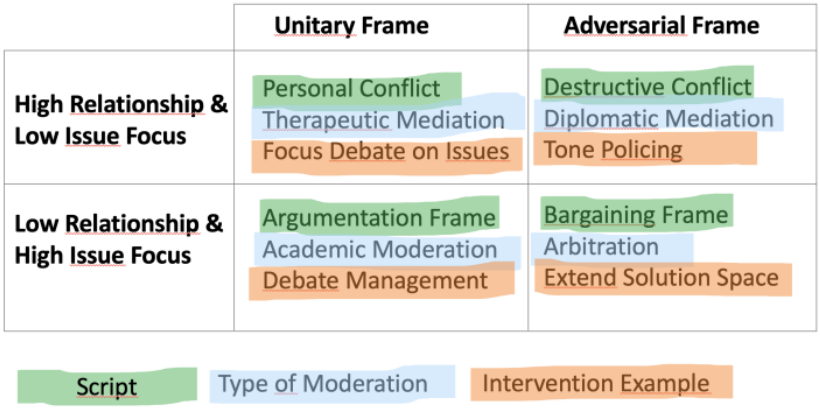

Figure 1: Moderation Context Model

The first dimension can be read as a dichotomy of a discussion being highly focused on a relationship and only little focused on the issues at hand; or the opposite. The second dimension is defined as follows: the unitary frame means that (given the topic, or in general) a consensus seeking disposition

can be assumed. In the adversarial frame individual interests are thought of as dominant and consensus is seen as secondary.

In the top left field of Fig 1, a consensus is thinkable but as the focus on the relationship is too high, the moderation intervention aims to bring back the issues. This is only true in social media where there is no point in fixing the relationship between mostly anonymous users. In a different setting, e.g. working environment, different strategies of managing conflict that take into account the relationship, the interests and the issues may be more appropriate

In the top right field, there is a personal conflict and participants enter with a adversarial frame of mind. Here, mediation interventions like tone policing or diplomatic statements may be adequate in order to reach a civilized level of debate.

The bottom left field focuses on the issues and a consensus seems possible. In this case the moderation needs to foster the argumentative content of the discourse by, for instance, weighting all arguments equally and drawing easy-to-follow inferences.

Finally, in the bottom right field, the focus is on the issues but individual interests overshadow a collective consensus. In this situation moderation should support the bargaining process. For example the Harvard negotiation model (Ury, 1993) or similar should be employed.

3.3.3 Moderation Strategies

Going through the different combinations of process functions of moderation, their connection to deliberation and their social context, five moderation strategies for user moderation in social media were distilled. Fig 2 summarizes our moderation concept.

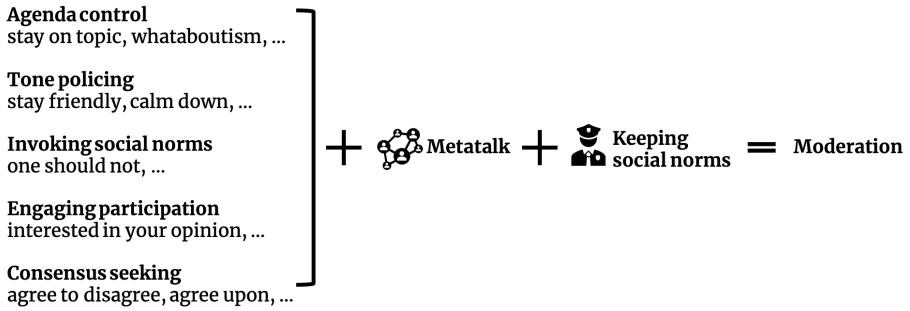

Figure 2: Moderation Concept

Moderation strategies aimed at agenda control involve guiding the direction and focus of the discussion. This can be achieved by controlling the agenda and ensuring the conversation centers around specific topics. Moderators may also

point out differing interest structures within the discussion, helping participants understand the underlying motivations and perspectives at play.

Tone policing strategies focus on maintaining a respectful and constructive tone within the discussion. Moderators may encourage participants to find common ground and seek compromise or end discussions peacefully. They may also intervene to improve relationships among participants, fostering an atmosphere of mutual understanding and empathy. By asking participants to explain their opinions and arguments, moderators facilitate comprehension of different perspectives.

Invoking social norms is another moderation strategy that relies on established community guidelines and norms. Moderators may highlight repetitive mistakes in the discussion, reminding participants to adhere to the agreed-upon rules. These interventions help maintain a harmonious and productive environment.

Engaging participation strategies aim to involve a wide range of perspectives and encourage active contribution from participants. Moderators may invite new perspectives, particularly in social media where attention mechanisms may limit exposure to diverse views. By broadening the range of voices and opinions, moderators foster a richer and more inclusive discussion.

Consensus-seeking strategies focus on building shared understanding and reaching agreement among participants. Moderators may support consensus building by summarizing important points and providing a synthesis of the discussion. They may also increase the bargaining space by breaking up polarized black-and-white thinking, encouraging nuanced and balanced viewpoints.

By employing these moderation strategies, discussions in social media communities can be effectively guided and facilitated. Agenda control ensures focused and relevant conversations, while tone policing promotes respectful dialogue. Invoking social norms maintains adherence to community guidelines, and engaging participation ensures diverse perspectives are represented. Finally, consensus-seeking strategies aim to build shared understanding and reach agreements.

After conducting initial field research, it was observed that norm enforcement and consensus seeking strategies were rarely observed in the targeted social platforms. As a result, we decided to prioritize the collection and classification of a large number of posts as "moderating" using the developed taxonomy, rather than attempting to capture the entirety of moderating behavior on those platforms.

3.3.4 A framework for describing the reactions to moderation

Drawing upon the documentary method and Mannheim's sociology of knowledge (Kleemann et al. (2013); Mannheim et al. (2022); Bohnsack (1999)), the subsequent framework was employed to elucidate the discourse patterns that emerge

subsequent to a partially moderating post. The ensuing categories represent the labels assigned to posts in a conversation.

1. Positive Counter-horizons:

- Elaboration: Expanding upon a point or idea to provide more details, examples, or explanations.
- Differentiation: Highlighting the differences between two or more concepts, ideas, or perspectives.
- Validation: Providing support or evidence to confirm the validity or correctness of a statement or argument.
- Ratification: Officially approving or confirming a decision, action, or agreement.

2. Negative Counter-horizons:

- Antithesis: Presenting a contrasting or opposing idea or concept.
- Opposition: Expressing disagreement or resistance to a particular viewpoint or proposition.
- Divergence: Discussing or exploring alternative paths or possibilities that deviate from the current direction or consensus.

3. Closing Thematic Conclusion:

- Positive Evaluation: Offering a positive assessment or judgment of the overall theme, topic, or argument.
- Opinion Synthesis: Bringing together different opinions, perspectives, or arguments to form a cohesive and balanced conclusion.

4. Ritual Conclusion:

- Change of topic: Shifting the discussion to a different subject or area of focus.
- Formal synthesis: Summarizing the main points or ideas discussed in a structured and organized manner.
- Rejecting the topic: Explicitly dismissing or refusing to engage with the current topic or proposition.

5. Emotive Reaction (no continuation or conclusion):

- Positive emotive/Humor: Expressing positive emotions or amusement through jokes, witty remarks, or lighthearted comments.
- Negative emotive/Sarcasm: Conveying negative emotions or disdain through irony, mockery, or caustic remarks.

6. Special case: Conclusion as metacommunication:

In this special case, the conclusion itself becomes a form of communication about the communication process. It may involve reflecting on the discussion, summarizing the main points, acknowledging different perspectives, or expressing a desire for further dialogue or exploration.

Naturally, with this perspective there won't be a perfect fit of textual data and the categories of interest. Sometimes discussion can be completely incomprehensible or very emotional. In the first case data points will be discarded as it is not possible to understand all kinds of communities on social media. In the second case these reactions will be labeled as emotional only if there is no rational category that could fit.

3.4 Hypotheses and Qualitative Research Questions

Based on the specified theoretical grounds, we derive at some specific questions that summarise our expectations. Our main research question revolves around the structure of reactions to the attempts of user moderation. As described earlier, we are dealing with social media and user responses towards intentions to shape communication through moderation means. One of the many responses is just to not respond but drop out of the communication; we cannot take responses for granted. On the other hand, users might also express their (dis)agreement with the intervention and continue the running communication. Moreover, other users might dig in to the communication – taking stances in favor or against the intervention. Our first hypothesis focuses on the response patterns and aims to extracting and analyzing the response patterns: What are (typical) response patterns of user reactions to moderation interventions? (**H1**)

Given there are reactions to the moderating post, it is crucial to differentiate between the various strategies employed and examine how they manifest in the discourse. Consequently, our second question delves into whether the reactions to the post align with the intended purpose of the moderating strategy (**H2**). For instance, if the moderation strategy was focused on maintaining topic control, it becomes imperative to analyze whether the communication actually got back on (topic) track. The same is true for the other two types of moderation: tone policing and elaboration support. With regard to H2, our specific expectations are:

H2a Agenda Control leads the communication back to the original topic.

H2b Emotion Control improves the "tone" of the discussion.

H2c Elaboration Support leads to more elaboration.

The third question is whether user moderation improves the deliberation quality of the following conversation as defined in the rational discourse framework (**H3**).

There is also some evidence that metacommunication like moderation leads to discussions about the metacommunication intervention. Hence, our last hypothesis states that moderating statements should invoke metacommunication (**H4**).

4 Data and Methods

4.1 Data Acquisition

As platforms Twitter and Reddit are considered. Most of the analysis were done on both Twitter and Reddit data. In the final sample only Twitter data was used. Both English and German posts were analyzed.

The discussions we are looking at are written exchanges that can be perceived as a conversation tree with the original post/tweet being the root node and answers-relations represented as the edges that lead to the answers that are the nodes in the tree. The leaves of the tree are answers that have no replies (retweets/quotes).

To find conversations that include moderation statements, in principle, three approaches could be applied. The first approach translates the task to a machine learning exercise: go through a randomly selected number of tweets and posts and label these accordingly – whether they contain a moderation statement or not. However, due to vast amounts of available data, finding moderation statements is extremely rare and this inductive approach is prone to fail.

The second approach builds upon apriori knowledge as to when conversations will most likely be moderated. This knowledge can be wrapped into computational procedures. For instance, if moderation is more likely to be seen when incivility increases, sentiment analysis might be used to decrease the amount of conversations to label. To succeed, this computational approach requires translating each moderation strategy into a (statistical) function. Even though this might work for some strategies, others are hard to proxy. Also, the amount of available tweets and posts is still very high.

The third approach – our applied approach – derives a dictionary of phrases that are most likely used when moderators intervene into a conversation. For each of the five moderation strategies and based on the predefined theoretical grounds of moderation, we employ a set of terms and phrases to reduce the amount of data. While on the one hand, these phrases should precisely define moderation –which speaks towards deriving phrases that are specific to certain tweets and posts– on the other hand, these phrases should show general applicability. In deriving the phrases, we have payed attention towards both of these goals. As an example of a phrase, think again about tone policing: if a tweet or post contains a phrase like "calm down" or "stay calm", this phrase might indicate that a user takes over a moderating role. It might also be the case that the exact phrase is used in a different context, not indication moderation. Hence, in a second

step, we go through the received tweets and posts and only mark moderation instances.

Based on the latter qualitative-deductive approach, we have deduced around 70 moderation phrases in both English and German[4]. These phrases were used as queries in Twitter and Reddit to download the conversation as well as the associated reply tree. To reconstruct conversational threads, we rely on our own Python library delab-trees[5]. In Fig 4, two different reply-trees are shown. Whereas the left tree is characterized by a root post that was almost exclusively answered by single replies, in the right tree, users more often respond to each other indicating some conversations have happened. Since we are mainly interested in conversations, we prioritize trees with longer chains of replies over those with a mushroom shape (one root post and many replies to this post, left tree in Fig 4).

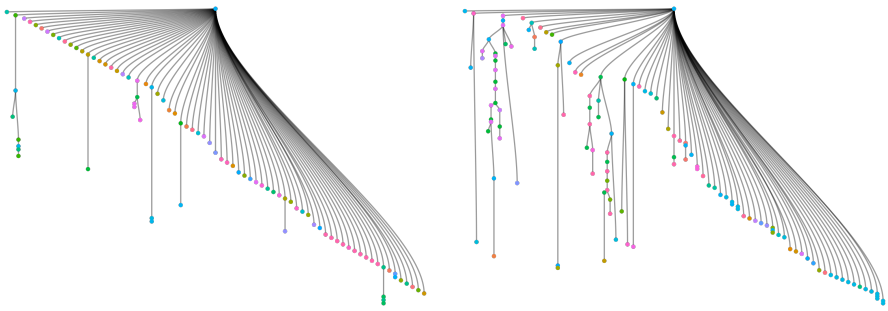

Figure 3: Two examples of conversation trees.

Additionally, since social media sites delete tweets and posts that are against their policies and are also required to delete content due to legal requests, some of the trees contain missing references. For instance, if a tweet gets deleted, the corresponding reply tree is broken. Hence, each of the conversation trees was validated. In the common case of deletion, the orphaned branches were attached to the root post. As the subsequent approach (see also next section) only takes into account reply chains excluding the root post this procedure does not lead to semantic errors.

To arrive at a large dataset with conversations containing moderation instances, two challenges remain: first, not all phrases are clearly indicating moderation; there is some ambiguity involved. Hence, we chose to label moderation instances based on a judgment made by three annotators. The three annotators had to judge whether a given tweet or post entails moderation. To provide some context, the annotators were given the root post, the two preceding posts, the post containing the queried phrase as well as the two following posts.

[4]The sheet with the phrases and their categories in terms of moderation strategies and type of social context will be attached in the appendix.

[5]https://pypi.org/project/delab-trees/

Overall, the inter-annotator agreement varies widely. While for some terms and phrases, no agreement could be made whether this tweet or post contains moderation, some other phrases show a high inter-annotator agreement of 0.94[6]. The disagreement is mainly due to difficulties in interpreting the degree of metatalk (see Fig. 2).

The second challenge is scaling: we just cannot manually label thousands (or even millions) of potential moderation candidates. Hence, for the purpose of this study, we chose to manually label 1.500 tweets and posts manually. We then query those phrases at a large scale that show both a high inter-annotator agreement as well as a high precision for moderation. Using this procedure, we end up with 8 phrases belonging to the strategies of agenda control, tone policing, and engaging participation. We finally arrive at a dataset of around 15.000 conversations with 4.6 million Tweets and Reddit posts. Overall, around 220.000 different users were involved in these conversations; with on average 17 users per conversation. On average, the longest reply-chain per conversation has a length of 7 tweets or posts. Since we are dealing with social media, most posts – especially tweets – are rather short: the average length of a post is 126 characters.

4.2 Reaction Pattern Labeling

To analyze the response patterns, we introduce the concept of conversation pathways or conversation flows. These are chains of replies that start with a root post (the first post with the opening statement of a conversation) and end in a leaf (the last contribution in a conversation). Using these structures as the unit of analysis facilitates the analysis process, because it maps the complicated structures to a quasi-linear conversation that can be interpreted analogously to a real life discussion. This is illustrated in Fig 4. Similar conversation modelling describes online conversations as either polyadic conversations (Magnani et al., 2012), reply graphs (Joglekar et al., 2020; Nishi et al., 2016; Cogan et al., 2012), or implicit thread structures (Wang et al., 2008).

It also allows the interpretation of a moderating post as an intervention. The result of the intervention can be computed as the sum of the effect on all conversation pathways that include the moderating post. Moreover, most annotating tools are not equipped to deal with tree structures. However, if transformed into conversation flows, these can be analyzed qualitatively like any other transcribed group discussions. Hence, also the usage of the documentary method which application domain originates from ethnographic field work with groups.

In order to analyze the effect of the moderating posts the context window of two posts before and two posts after the moderation intervention is selected. The size of the context window ($n = 5$) is a compromise between the size of limited available flows of size greater than n and the need to keep the amount of text for the qualitative analysis limited to a manageable amount. At this stage

[6]The inter-annotator agreement is based on Krippendorff's alpha (Krippendorff, 2004).

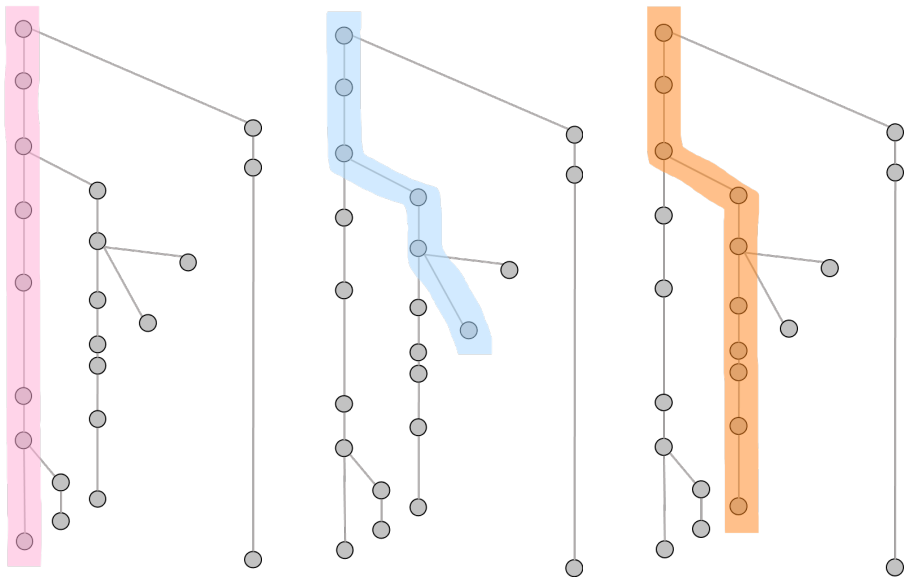

Figure 4: Conversation flows within the reply tree: if the trees are viewed as as list of branching linear conversations it is illustrated in pink the first flow, in blue a second and in orange a third. More flows exist in this tree.

there were 4971 samples with the post being labeled as partially moderating. This is due to a second dataset being added in addition to the labeled tweets in order to increase the size of the dataset from which to draw the rare deep conversation flows. Here, the moderating tweets are labeled automatically based on the phrases that had a high intercoder-reliability during the labeling of the main corpus.

From the combined dataset of the reply trees 2740 conversation flows were sampled that include a moderating post. From these there are 1948 that include the predefined length of a 5 context-window. From these there are 1540 5-windows with no self-answers.

In order to simplify the analysis further the 1540 5-windows were grouped based on the moderating post which can be contained in several flows. From each group only one flow instance was selected in order to avoid confusion or bias during the labeling process. Otherwise, it would have been important to consider additional factors in the analysis, such as the inclusion of mentions and other multiple references, along with the timing of the answers. After this process and removing problematic flows (answers in different languages for example) the final dataset consisted of 508 5-context windows. In Table 1, the number of flows is counted for each of the three sampled moderation strategies. Elaboration support is by far the most often sampled strategy, followed by emotion control and agenda control.

Moderation Strategy	Count
Elaboration Support	342
Emotion Control	122
Agenda Control	44

Table 1: Moderation Strategies and Count of 5-context windows

After sampling the $n = 5$ context windows the moderating posts and the following two posts were exported alongside the symbolic pattern of authors. For example with the pattern "*abcba*" the author c wrote the moderating post answering b who answered in return followed by a again.

This was intended to give some insight into the dynamics of the conversations without blowing the workload of the labeling task out of proportion. The fourth and the fifth post were labeled according to the rational discourse framework discussed in the theory section. Moreover, as a second verification step, the moderating posts were relabeled as moderating or not. Only very few data points had to be removed due to post not being moderating. This, again, demonstrates the general reliability of the chosen sampling procedure.

5 Results

5.1 Engendering Answers

To analyze (typical) response patterns (H1), we first calculate the probability of seeing an answer to a moderation intervention. Assuming to self-answers, the probability is 0.79. As can be seen from Table 2, most of the user moderation takes place in the first two thirds of the conversation flows which means that there is a tendency for the conversations to go on longer if user moderation has taken place.

Position in Conversation Flow	Probability of Occurance
1/3	0.43
2/3	0.36
3/3	0.22

Table 2: Probability of user moderations

These observations may be influenced by the context window itself. Many conversation flows do not span the length of five consecutive posts. If the position was computed by looking without this limitation this would increase the size of the first and third position in the conversation flow. Moreover, by definition, moderation was not labeled if the post was the original post of the conversation. This would also influence the distribution of the positions.

An interesting observation is that most longer flows are dialogues which can be

seen in the following list of patterns and their normalized value counts. Patterns with less than 1% incidence were omitted. Figure 5 shows the different patterns. In quite a few cases authors answer their own posts. Even if these cases were eliminated, the dialogue pattern (*ababa*) is by far the most dominant one.

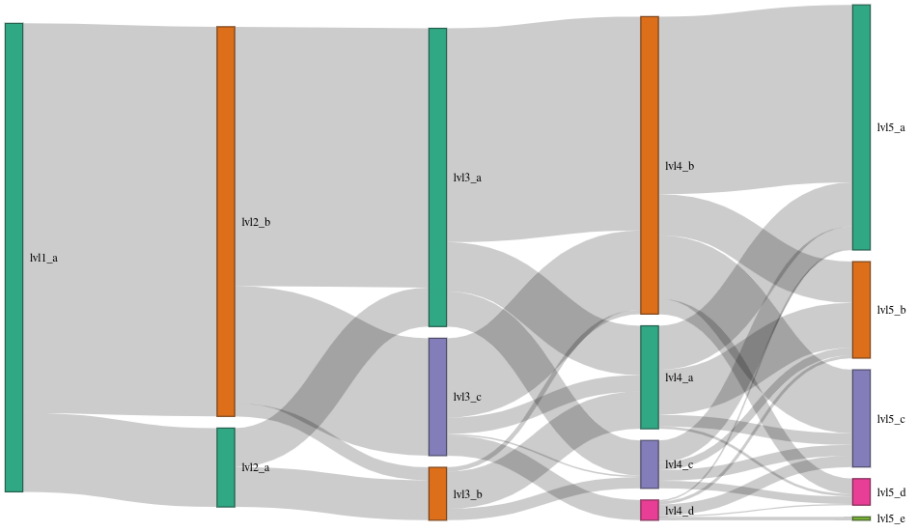

Figure 5: Patterns of repeated user posts in the 5-post context window

It is more likely for the window to have three authors or less than it is to have more than three. Only 3% of cases include 5 authors. This changes the perception of the conversation flows as group discussions of large groups. They resemble more of a small table of people.

5.2 The effectiveness of the different strategies

The answer to H2 must be differentiated between the different strategies analyzed. Figure 6 shows the probabilities of a certain pattern.

Regarding Agenda Control, the second hypothesis (H2a) states that moderation should lead the communication back to the original topic. However, as given in Fig 6, the reactions to Agenda Control are found to be mixed. This implies that while emotional reactions may not be the dominant response, there is still variability in the way individuals react to Agenda Control. To some degree, agenda control results in fewer divergences and oppositions compared to other reactions. Hence, it seems to effectively fulfill its function at a discourse level, influencing the direction and focus of the discussion. However, a comprehensive analysis of whether the discussion adheres to a given topic content-wise would require an alternative methodology, such as natural language-based topic modeling.

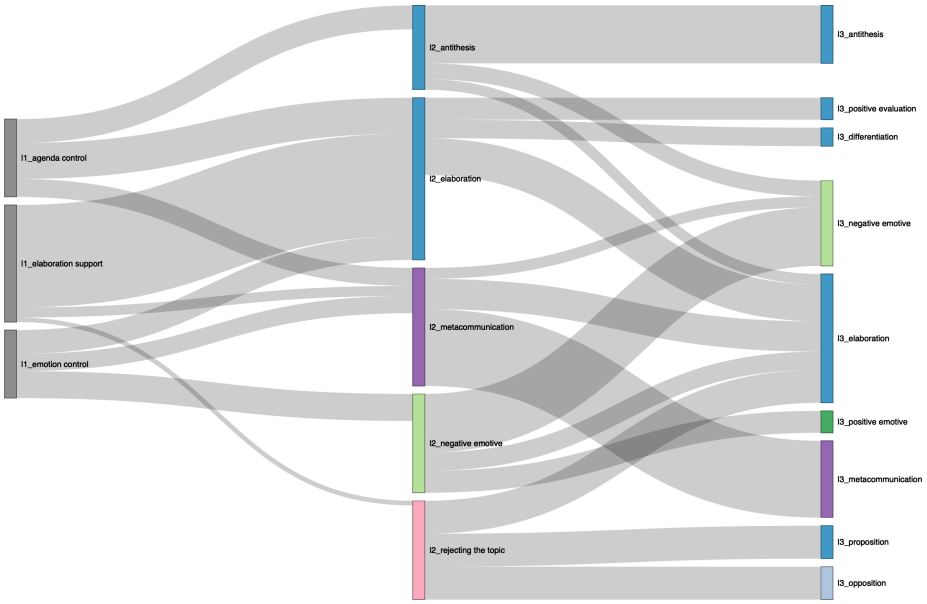

Figure 6: Reactions to user moderation in terms of discoursive structure.

Moving on to Emotion Control, H2b states that it leads to an improved tone. However, negative emotive reactions are more likely to occur, along with an increased likelihood of conversation concluding posts. This suggests that attempts to control emotions may not effectively improve the overall tone of the discussion. Instead, they may inadvertently contribute to negative emotional reactions and potentially lead to the conclusion of conversations.

Regarding Elaboration Support, H2c posits that it leads to a substantial increase in deliberation. This indicates that when users provide elaborated responses and support their arguments or statements, it encourages more in-depth discussion and consideration of different perspectives. In combination with H3, H2c suggests that due to the high incidence of elaboration support, emotive reactions are relatively infrequent. However, it is worth noting that a significant portion of elaboration support occurs within contexts that involve sharing information rather than engaging in arguments or discussions. Furthermore, approximately 10 percent of the instances labeled as elaboration support were additionally classified as professional support. These instances raise questions about whether they should be grouped under the umbrella of user moderation or deliberation since they may differ in their nature or purpose.

In summary, the findings related to Agenda Control, Emotion Control, and Elaboration Support provide insights into their effects on the discourse. The results suggest mixed reactions to Agenda Control, potential drawbacks of Emotion Control in terms of tone improvement, and the significant impact of Elabo-

ration Support on promoting deliberation. However, further analysis and consideration are necessary, especially regarding the categorization of certain instances within these strategies.

5.3 The influence on deliberative talk

In order to generalize the effect of user moderation to the effect on deliberative quality (H3) the positive horizons (elaboration, differentiation and validation) and the productive conclusions (opinion synthesis and positive evaluation) were marked deliberative and the rest as non-deliberative. A special case are positive emotive reactions that were grouped under deliberative. From a strict rationality-oriented definition even positive emotions should not be considered as a productive continuation of the argument. However, most users on social media are humans.

Figure 7 shows that all user moderation strategies engender deliberative talk. However, emotion control seems to have less of an effect. It is also interesting that once deliberative talk has started it remains deliberative whilst non-deliberative talk can turn to deliberative. This may be an artifact of the high incidence of dialogues.

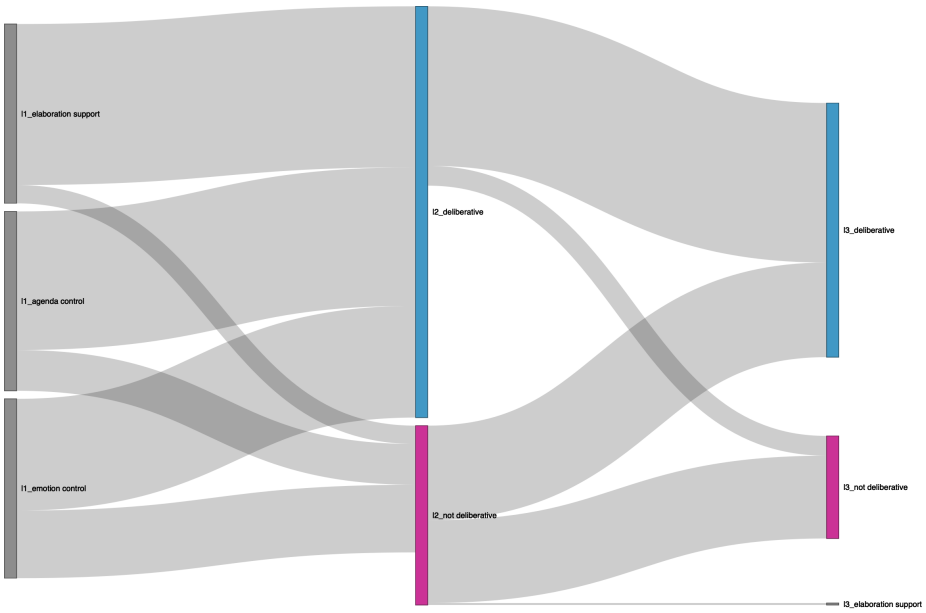

Figure 7: Moderation Strategies and Deliberation

Most 5-context window patterns are "*ababa*". This means that the second reaction is often the user who wrote the moderating post. One interpretation of this finding is that the user who wrote the partial moderation is trying a second time to improve the discussion. In terms of participation this should

be viewed as a negative finding in that moderating posts do not transform the dialogues to group discussions. However, this also is plausible as the strategy that invites participation was not part of the strategies being investigated. For these questions, the full set of moderating strategies would have to be tested on the full discussion and not the context windows alone. For these reasons not too much store should be set on the finding that user moderation produces deliberative reactions. It should be viewed as an indicator and not a causal statement. Another explanation is that elaboration support produces longer conversations and that the sample might be skewed towards deliberative reactions for that reason.

In the case of elaboration support a small portion of reactions contains more elaboration support. This could be considered deliberative. However, we chose to look at these cases as an exception were the unit of analysis should be a sequence of moderating posts and their answers rather than a single post. Distributing user moderation over several posts spread in the conversation was not in the scope of the analysis, as it would require to read the entire conversations for the raters. An interesting facet of this pattern is that it only occurs when the first reaction to the user moderation was not deliberative.

5.4 The influence on metacommunication

Figure 8 shows that metacommunication does bring out more metacommunication (H4) but it also produces a positive effect on deliberative quality. Positive counter-horizons are very likely to be the reaction to metacommunication. In comparatively fewer cases negative emotive reactions to the metacommunication ensue.

If moderation is seen as metacommunication, it would suggests that metacommunication does not always lead to increased metacommunication. However, the better interpretation is that moderation is an exception and produces different outcomes than metacommunication in general.

Metacommunication is defined in the discoursive analysis framework as a special case of a conclusion. This definition should be extended with user moderation as an horizon-opening version of metacommunication. This fits in nicely with the framework and should be considered a theoretical contribution.

6 Discussion

Overall, we demonstrate the potential of user moderations: When users intervene in a discussion, it certainly has some effects. Based on our approach, however, we cannot conclude whether the strategies themselves are effective; this would require a more in-depth qualitative approach. The reactions are highly individualized and detecting patterns of divergence or consensus by computational means remains a difficult task. Moreover, our research shows some clear limitations: First, only three of five strategies could be analyzed successfully, because of the

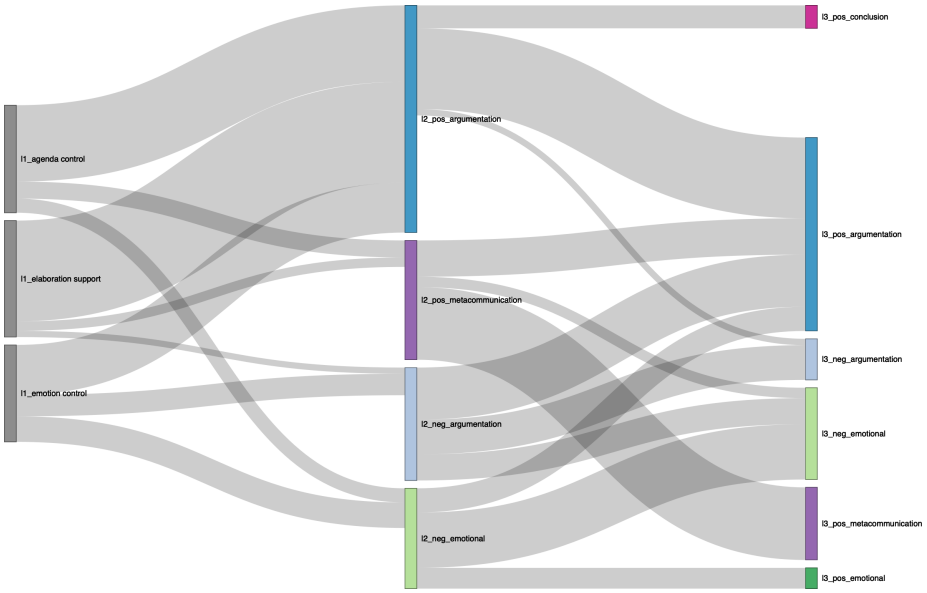

Figure 8: The effect of metacommunication

low scores in intercoder reliability during the sampling. The sampling was very difficult and reliability scores were low for many phrases and strategies. With a random sample from social media it is not easy to understand the context fully given a context-window of only a couple of posts. Since the amount of user moderation is so little, limiting the sample to political topics like the originally intended was not feasible. This also created some noise in terms of usage of social media. Some data had to be discarded because it resembled support groups (for trading, technical expertise etc.). In other cases the topics were so specific (i.e. gaming, African Politics) that the confidence of the raters was low. In the final sample the different strategies were not equally represented. As the analysis is based on single classes the skewed sample does not change any conclusions. However, the low data count for agenda control should be considered an encouragement to investigate how the results might change with higher numbers of data points.

Second, using phrases or language features in general provides some noise in the pipeline early on. Despite the human-in-the-loop approach some uncertainty exists whether "back to topic" sometimes is a self-reference (I should get back to topic) or a suggestion (We should get back to topic) or normative statement (ONE should get back to topic). However, with the second validation step, we show that our sampling procedure mostly extracts only moderation instances.

The labeling was done for English and German. However the results were aggregated without investigating the effects of the language at hand. Since the

phrase filtering and other aspects of the research design are highly dependent on the language, this is an important next step for further research using the corpus.

Since the users' primary goal is not moderation or improving deliberation, interpretation of when a post classifies as moderating was difficult. Using a crowd-based approach for generating moderating posts could be a good way of testing the stability of the results in this regard.

Third, limiting the unit of analysis to the 5-post context window allows for manual labeling but also creates some uncertainty whether the essence of the conversation was captured. Computational measures that analyze the complete conversation flows such as sentiment analysis or topic labeling could be employed in addition to the pattern analysis.

Another approach may have been to focus on the psychological prerequisites of rational discourse. For instance, if one assumes that an angry mob cannot argue rationally, it would be sufficient to measure negative emotions like hate, fear and anger which might be easier than testing the discourse for rationality.

Overall, the novelty of the findings in this study is considerable, given the approach employed. Unlike many other approaches that primarily focus on mitigating hate speech or reducing its impact, this study tackles the intricacies of conversation dynamics without assuming the presence of aggressive conflict. While this approach presents a challenge in terms of identifying clear-cut examples (as they are not as overt), it also increases the likelihood of uncovering instances of change. By studying conversations where participants are not heavily influenced by heightened emotions or polarized views, the potential for meaningful shifts in discourse becomes more apparent.

7 Conclusion

Our theoretical contribution is to place moderation within deliberation theory and script theory. From a empirical perspective, computational methods of data mining and data modelling of the conversation flows were used to provide the data for qualitative analysis. Using pattern analysis of behaviour or cultural scripts that guide the reactions to user moderation provided a basis for measuring the impact on deliberative quality of user moderation without judging the conversation by normative standards such as the quality of the consensus seeking or the level of participation and openness.

The results show a trend that even partial user moderation is usually successful in terms of continuation of the discussion. As an unintended result it can be observed that in the platforms investigated (Twitter and Reddit) the dialogue pattern is very typical which means that most of the 5-context windows only contain three different authors and the pattern *ababa* has the highest incidence. Initially the results were disappointing when looking for fully moderating posts

but with the information that most of the conversations are small table styled it explains why larger references such as "we as a group should do this and that" are rarely found. In a dialogue group norms are not enforced but changed to personal attacks (i.e. you should not be/do this or that, instead of one should not).

Elaboration support seems to be easier and more reliable than agenda control or emotion control. This is plausible as the phrase alone stipulates a resource-oriented communication rather than enforcing conversational rules. Although it is not part of the sample, the assumption would be that value enforcement such as tolerance would result in even less deliberative reactions. This should be the focus of subsequent studies.

Supporting consensus-seeking remains the gold standard for both moderation and deliberation. Here, the reality of social media communication seems to be that consensus-seeking is not the priority on users' mind. Despite reading and labeling many posts in different platforms on social media, this kind of behaviour was hardly found using the phrase based data acquisition strategy.

In order to complete the picture on different strategies these infrequent strategies should be researched by producing examples artificially. These kinds of moderating posts could be created crowd-based (using Mechanical Turks or similar). The definition of user moderation, the developed methodology and the corpus allow for further inquiries into the effectiveness of different moderation strategies and their constraints that depend on social and cultural scripts.

The approach taken here has the advantage that it is non-reactive (in the wild) and tackling a hitertho little researched topic. The methodology and the created corpus can be reused for similar research questions and will be published as open source (links forthcoming).

The results can be viewed as a good sign that non-restrictive user authored moderation can succeed in improving the deliberative quality of online discussions without relying on deletion or bans. Most of the time user moderation engenders more discussion and reactions can be classified as deliberative more often than as destructive.

Acknowledgement

This work was funded by the Volkswagen Foundation under grant 98 540 "Deliberation Laboratory (DeLab)". The authors would like to thank Pauline von der Haar and Fabian Lochner for their help.

8 Bibliography

Black, L. (2013). Framing Democracy and Conflict Through Storytelling in Deliberative Groups. *Journal of Deliberative Democracy 9*(1).

Bohnsack, R. (1999). Dokumentarische Methode. In R. Bohnsack (Ed.), *Rekonstruktive Sozialforschung: Einführung in Methodologie und Praxis qualitativer Forschung*, pp. 34–80. Wiesbaden: VS Verlag für Sozialwissenschaften.

Bormann, M., D. Heinbach, and M. Ziegele (2021). " can we please stop yelling at each other just because it's the internet?" comparing incivility perceptions of community managers, users, and ac-tivists in online comment sections. In *Weizenbaum Conference 2021: Democracy in Flux–Order, Dynamics and Voices in Digital Public Spheres*, pp. 5. DEU.

Bächtiger, A. and J. Parkinson (2019). *Mapping and Measuring Deliberation*. Oxford University Press.

Cai, J. and D. Y. Wohn (2019). What Are Effective Strategies of Handling Harassment on Twitch? Users' Perspectives. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, CSCW '19, New York, NY, USA, pp. 166–170. Association for Computing Machinery.

Chambers, S. (2003, January). Deliberative democratic theory. *Annual Review of Political Science 6*(1), 307–326.

Cialdini, R. B. (2003, August). Crafting Normative Messages to Protect the Environment. *Current Directions in Psychological Science 12*(4), 105–109. Publisher: SAGE Publications Inc.

Cogan, P., M. Andrews, M. Bradonjic, W. S. Kennedy, A. Sala, and G. Tucci (2012). Reconstruction and analysis of Twitter conversation graphs. In *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research - HotSocial '12*, Beijing, China, pp. 25–31. ACM Press.

Edwards, A. R. (2002). The moderator as an emerging democratic intermediary: The role of the moderator in Internet discussions about public issues. *Information Polity 7*(1), 3–20.

Friess, D. (2021). Collective Civic Engagement and Civic Counter Publics: Theoretical reflections upon a new phenomenon. In *Weizenbaum Conference 2021: Democracy in Flux–Order, Dynamics and Voices in Digital Public Spheres*, pp. 7. DEU.

Grimmelmann, J. (2015). The Virtues of Moderation. *Cornell Law Faculty Publications 42*(17), 43–109.

Habermas, J. (1981a). *Theorie des kommunikativen Handelns: Band 1*. Frankfurt am Main: Suhrkamp.

Habermas, J. (1981b). *Theorie des kommunikativen Handelns: Band 2*. Frankfurt am Main: Suhrkamp.

Habermas, J. (1982). *Strukturwandel der Öffentlichkeit*. Frankfurt am Main: Suhrkamp.

Habermas, J. (2009). *Diskursethik*. Frankfurt am Main: Suhrkamp.

Han, S.-H., L. M. Brazeal, and N. Pennington (2018). Is civility contagious? examining the impact of modeling in online political discussions. *Social Media+ Society 4*(3), 2056305118793404.

Joglekar, S., S. Velupillai, R. Dutta, and N. Sastry (2020, July). Analysing Meso and Macro conversation structures in an online suicide support forum. Technical Report arXiv:2007.10159, arXiv. arXiv:2007.10159 [cs] type: article.

Kleemann, F., U. Krähnke, and I. Matuschek (2013). Dokumentarische Methode. In *Interpretative Sozialforschung*, pp. 153–195. Wiesbaden: Springer Fachmedien Wiesbaden.

Kluck, J. P. and N. C. Krämer (2022). Appraising uncivil comments in online political discussions: How do preceding incivility and senders' stance affect the processing of an uncivil comment? *Communication Research 40*(4), 453 – 479.

Kraut, R., P. Resnick, S. Kiesler, M. Burke, and Y. Chen (2012). *Building Successful Online Communities: Evidence-Based Social Design*. The MIT Press. MIT Press.

Krippendorff, K. (2004). Reliability in content analysis. *Human communication research 30*(3), 411–433.

Magnani, M., D. Montesi, and L. Rossi (2012, June). Conversation retrieval for microblogging sites. *Information Retrieval 15*(3), 354–372.

Mannheim, K., D. Kettler, V. Meja, and N. Stehr (2022). *Strukturen des Denkens* (3. Auflage ed.). Suhrkamp-Taschenbuch Wissenschaft. Frankfurt am Main: Suhrkamp.

Molina, R. G. and F. J. Jennings (2018a). The Role of Civility and Metacommunication in Facebook Discussions. *Communication Studies 69*(1), 42–66.

Molina, R. G. and F. J. Jennings (2018b). The role of civility and metacommunication in facebook discussions. *Communication studies 69*(1), 42–66.

Nishi, R., T. Takaguchi, K. Oka, T. Maehara, M. Toyoda, K.-i. Kawarabayashi, and N. Masuda (2016, May). Reply trees in Twitter: data analysis and branching process models. *Social Network Analysis and Mining 6*(1), 26.

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New media & society 6*(2), 259–283. Publisher: Sage Publications.

Porten-Cheé, P., M. Kunst, and M. Emmer (2020). Online civic intervention: A new form of political participation under conditions of a disruptive online discourse. *International Journal of Communication 14*, 21.

Pradel, F., J. Zilinsky, S. Kosmidis, and Y. Theocharis (2022). Do users ever draw a line? offensiveness and content moderation preferences on social media.

Rosenberg, M. B. (2016). *Gewaltfreie Kommunikation: Eine Sprache des Lebens*. Junfermann Verlag GmbH.

Seering, J. (2020). Reconsidering community self-moderation: the role of research in supporting community-based models for online content moderation. *Proc. ACM Hum.-Comput. Interact 3*, 107:28.

Sternberg, J. (2012). *Misbehavior in Cyber Places: The Regulation of Online Conduct in Virtual Communities on the Internet*. Lanham, Maryland: Rowman & Littlefield.

Stromer-Galley, J. (2007, January). Measuring deliberation's content: A coding scheme.

Tucker, J. A., A. Guess, P. Barberá, C. Vaccari, A. Siegel, S. Sanovich, D. Stukal, and B. Nyhan (2018). Social media, political polarization, and political disinformation: A review of the scientific literature.

Ury, W. (1993). *Getting past no: negotiating your way from confrontation to cooperation*. Bantam Books.

Wang, Y.-C., M. J. M. Joshi, and W. Cohen (2008). Recovering Implicit Thread Structure in Newsgroup Style Conversations. *Proceedings of the International AAAI Conference on Web and Social Media 2*(1), 152–160. Number: 1.

Watson, B. R., Z. Peng, and S. C. Lewis (2019). Who will intervene to save news comments? deviance and social control in communities of news commenters. *New media & society 21*(8), 1840–1858.

Wohn, D. Y. (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*, CHI '19, New York, NY, USA, pp. 1–13. Association for Computing Machinery.