Strategies for Deliberative Moderation in Social Media

Abstract

This study delves into the intersection of moderation and deliberation theories by investigating moderation strategies employed by common users on social media platforms to encourage deliberation. This paper seeks to synthesize theoretical perspectives into a model of deliberative moderation. It also addresses a pressing practical question whether the developed model can be used to identify strategies of deliberative moderation in online settings employed by ordinary users. Through a content analysis, we find that the strategies *emotion control*, *agenda control* and *elaboration support* can be reliably classified and found on social media using this new model.

Keywords

Social Media, Deliberative Democracy, Moderation Strategies, Content Analysis

1. Introduction

Discussions in social media on political matters could be viewed as deliberative as they are relatively power-free and open for participation. However, this idealistic view of the public sphere has been challenged, and the many reports of hate speech, automated trolls and short attention span present a dire picture. Here, moderation might be a possible antidote that could improve deliberative quality.

The literature has not been clear on what constitutes a *discussion* in social media technically and conceptually, so as to meet the criteria for being classified as a deliberation. Justification (Graham 2003, Gold et al. 2017), Reciprocity (Graham 2003), Respect (Papacharissi 2004, Rowe 2015, Ziegele 2022) and other aspects of deliberative talk (Smith 2012, Stromer-Galley 2007) only make sense analyzing if there is an exchange of views that lasts more than one post-and-reply dyad. In contrast to civility (Adler 2005, Coe 2014, Papacharissi 2004) which can be observed within a single post, the concept of deliberation derives from the image of longer exchanges. Applying the adjective *deliberative* to a solitary speech act has muddied the waters, when discussing the online public sphere, too.

In a similar vein content moderation has been limited to using posts as the unit of analysis, when discussing deplatforming, deleting posts or banning users. Single destructive comments can have an effect on the following discussion but with the analysis of counter-speech, journalist interventions (Stroud et al. 2015) or collective civic moderation (Friess et al. 2022) the focus is to research what prevents deliberative discussions instead of what promotes them.

Moderation understood as deletion or flagging is omitted here for another reason: we are constructing an analogy to connect to previous research on real-life deliberation. This allows the developed framework for deliberative moderation to be applied in other contexts than social media. Given this direction of research, deletion is not an option as you could not erase someone's speech act in real life.

This paper fills both above mentioned gaps, first the assumption is checked whether there are processes in social media that qualify as equivalent to a real-world discussion and if yes, these are analyzed for moderating behavior of ordinary users. For this purpose the following research questions are explored: (RQ1) How can a theoretical framework for deliberative moderation be conceptualized to suit diverse application contexts? (RQ2) How can different interpretations of deliberation theory be applied to understand the nature and dynamics of longer reply threads in social media discussions, and what theoretical implications does this have for conceptualizing deliberation as a process in digital environments? (RQ3) How can the developed framework for deliberative moderation be applied to identify and classify deliberative moderation in social media processes, and what criteria does the framework use to differentiate between varying degrees or types of deliberative moderation?

The concept of deliberative moderation is operationalized using these steps: first, different functions of moderation are considered based on the literature and their corresponding embedding within deliberation. Limiting the functions of interest to the process function, different social contexts are examined regarding what moderation looks like. The different social contexts of moderation are then modeled according to the focus and the dispositions of the participants in the conversation. Next, social media is considered as an application context for deliberative moderation. After this, the deliberative functions of moderation, the process aspects and the developed model of social contexts are integrated. These strategies are represented by phrases that are used to filter the data available in social media to find candidates for user moderation. In the following case study these categories are used to code the material leading to a secondary inductive category building process (Meyring 2004; Kukartz 2016; Ruin 2017). The main goal of the case study is to show that the framework can be applied to categorize different deliberative situations and corresponding moderation functions.

2. Related Research

Growing divergence becomes most obvious in social media, where different groups, attitudes and political leanings meet, and interact freely. With growing disagreement, the necessity of tolerance toward the positions of others becomes ever more important (Grube et al., 1994). Education is only a long term solution (Darmody, 2016; Rapp & Freitag, 2015; Rombout et al., 2022; Türkkahraman, 2014). There is ample research on countering hate speech in general on social media (Alsagheer, 2022; Ganesh, 2020; Yildirim, 2021). In terms of content moderation (Watson 2019; Seering 2020; Pradel 2022; Porten-Cheé 2022), one widely researched intervention is the use of counter-speech, which can be created by platform operators, through citizen moderation (Ziegele & Jost, 2020; Löb & Wessler 2021), or computer-automated means (Doganc, 2023; Bonaldi, 2022). Given the sizable amount of incivil speech shared online, scalable AI-based natural language generation has been suggested by several studies (Tekiroglu, 2020; Chung, 2021), even though being limited by the production of generic and repetitive responses to hate speech (Doganc, 2023).

In our research, we chose to focus on in-the-wild user moderation rather than relying on existing networks of civic citizens aiming to enhance the social media landscape. This made the sampling process harder but was conceptually necessary: firstly, access to these established networks is often limited or restricted, rendering them less feasible for widespread and inclusive moderation. Moreover, there's a prevailing concern about neutrality (Spada & Vreeland 2013); while these civic groups possess commendable intentions, they sometimes exhibit biases or leanings that can skew the moderation process. Notably, many of these networks prioritize anti-hate speech measures, which, though undeniably crucial, can sometimes overshadow the broader goal of fostering consensus-seeking and nurturing productive, rational discussions.

There are existing studies on moderation and deliberation that also deal with online environments (Muhlberger 2005, Epstein & Leshed 2016). This research continues that tradition but takes on three

new challenges: (1) study deliberative practice in the wild instead of a simulated environment or one that is purpose-built, (2) challenge the assumption of what constitutes a deliberative discussion in social media (3) not assuming an explicit moderator role or a specific user group.

3. Defining User Moderation as a Deliberative Act

According to Grimmelmann (2015: 47), moderation can be defined as 'the governance mechanism that structures participation in a community to facilitate cooperation and prevent abuse'. The idea 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.

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 that 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 comprehensible and constructive (Graham, 2003), 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 and does not imply institutional bonds like a teacher role.

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. 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. As Mutz (2008) points out, deliberation is always a social situation and 'political reasoning is often motivated by goals other than accuracy'. For this reason – and based on the definition by Edwards (2002) – **partial deliberative moderation** is defined *as the sentence within a post or tweet that appeals to a social norm that tries to structure participation to foster cooperation within the discussion process itself as a collective, purposeful activity that fits the social context of the situation*.

4. What does Moderation mean in different contexts?

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 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. In a setting with strong interests such as the United Nations Assembly or a trade dispute, the cooperative mode would be bargaining and trying to find an optimal compromise (Fisher & Roger & Patton 2015). Finally, in a couples' therapy the cooperative mode would be to discuss the relationship. Here, it might be appropriate to bring in all the emotions involved. For these reasons, it is important to classify a situation to apply the right type of communication and moderation theory (Bandler & Dilts 2015).

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 (Mansbridge 1980; Karpowitz & Mansbridge 2005), on the other hand conflicts may be about relationships or issues. In Fig 1, using these combinations of dispositions as dimensions, we derive a context model for moderation.

	Unitary Frame	*Adversarial Frame*
High Relationship & Low Issue Focus	Personal Conflict Therapeutic Mediation Focus Debate on Issues	Destructive Conflict Diplomatic Mediation Tone Policing
Low Relationship & High Issue Focus	Argumentation Frame Academic Moderation Debate Management	Bargaining Frame Arbitration Extend Solution Space

Legend: Ongoing Social Script, Type of Moderation, Intervention Example

Figure 1: Moderation Context Model (MCM)

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 mainly 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 consider the relationship or the underlying interests may be more appropriate. In the top right field, there is a personal conflict and participants enter with an adversarial frame of mind. Here, mediation interventions like tone policing or diplomatic statements may be adequate to reach a civilized level of debate.

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

Finally, the bottom right field captures a situation where 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.

The Moderation Context Model (MCM) and the definition of deliberative moderation come together mainly in the bottom left corner of the context model. Indeed, it has been a common criticism of Kant, and later Habermas that their ideals were based on the image of an academic discussion which is issue-focussed and assumes a unitary concept of a man, and that arguments have a certain power. The main point here is not to discuss these assumptions but to point out that they fit perfectly into the MCM as

one of the quadrants. The bottom left can be called the ideal quadrant, and perfect deliberative moderation would take place there as it is the one with the highest issue focus, the most consensus-seeking and common goods thinking. Even in the ideal quadrant with a high focus on the issues and a general disposition to find common ground, deliberative moderation might be necessary: it could be the case that too few perspectives are represented (another classical deliberative motive) or the discussion could be an instance of a echo chambers (Flaxman & Goel & Rao 2016). In this case, diverging opinions would be suppressed as a result of the social dynamics of echo chambers. This would force the moderator to play the role of a devil's advocate or invite others to do so. This is an important special case as it highlights the difference between this concept of deliberative moderation and countering incivil speech: in a solidified mono-group deliberative quality would lack because of redundancy, intolerance of new perspectives, or lack of arguments. But it is perfectly imaginable that all this happens in a polite, respectful tone.

Another special case is the area of fake news or differing belief systems. This can be viewed as the biggest hurdle of deliberative talk, because arguments lose their impact if the underlying assumptions are in question. Nevertheless, it should be noted that a functioning discursive space should be able to mitigate some of the negative effects of misinformation, (Garimella et al., 2018, p. 9) as [rational] "discussion [...] brings out facts". In this case the moderator would have to foster the reflection of belief systems by engaging in Socratic dialogue which means delving into the assumptions by asking critical questions. Beliefs can be thought of as a third dimension where in the origin (bottom left of the MCM) common beliefs and agreement on facts exists. The more adversarial the positions become or the more the issue-focus is lost the less agreement on reliable sources of knowledge there is. By mapping out these assumptions on a continuous plane, the MCM provides a tool for the analysis of real-world deliberative moderation strategies: moderating intervention can be inspected regarding their intent of steering a discussion towards a more issue-focused tone, all the while assuming that

participants want to find common ground or that the moderator thinks that they can be convinced to do so. Furthermore, irrational, adversarial discussions can be placed in the model (top right), too. In this circumstance, appropriate moderation strategies can be thought of that apply to the realities of non-academic flaming. The MCM covers the main literature fields on moderation (diplomacy, academic moderation, mediation and arbitration) but also connects well with political theory as the fundamental distinction between unitary democracy and adversarial democracy is included as one dimension. It also covers the typical questions of rationality vs. emotions with online deliberation: i.e. Kim (2016) on anger influencing information retention. The MCM only claims to be a theoretical construct that streamlines research perspectives. Further research is needed to answer the question whether it covers all possible situations for deliberative moderation. This study tries to show the model's merit by applying it to the complex field of deliberative moderation in social media.

5. (Deliberative) Group Discussions in Social Media Reply-Trees

The emergence of social media platforms introduced a new dimension to the study of deliberation theory. Initially, the internet was seen as a realm of relative freedom of speech and a potential catalyst for fostering international publics. However, the rise of hate speech and the proliferation of artificial destructive agents have cast these optimistic notions in a more critical light (Tucker et al. 2018: 32). Amidst these concerns, some crucial aspects of deliberation theory, such as rationality, productive dialogue, thematic focus, and consensus building, have suffered. These elements are essential for the proper functioning of the public sphere as a channel for social progress.

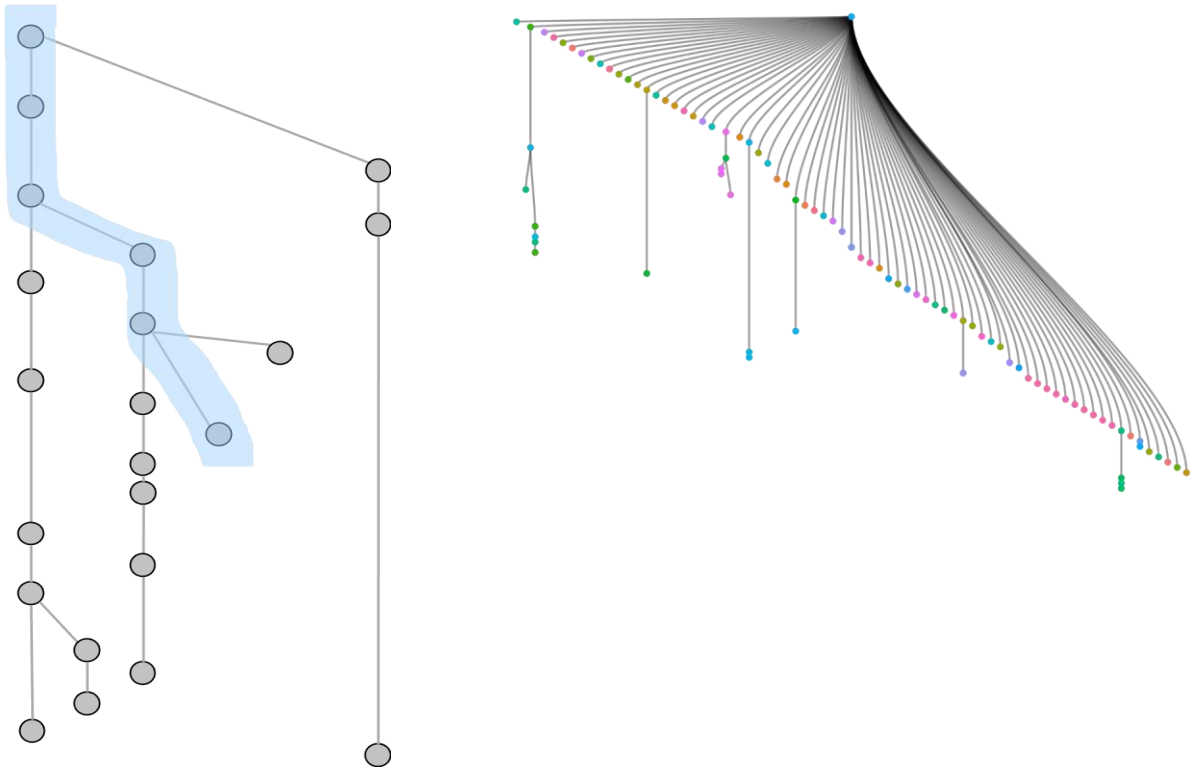
Not all communication on social media platforms embodies deliberative ideals. However, we contend that certain discussions within social media constitute a discursive form of public space. To undertake this exploration, we assume that certain aspects of social media can be interpreted as both rational and generally accessible, even if not universally so.

Delving into the works of Habermas, we find a tension between the preservation of rationality and the guarantee of general accessibility within a public sphere (Habermas 1982: 346). While rationality is a demanding expectation, transforming communication on social media to a level of reasonableness should be attainable. Habermas, in his Theory of Communicative Action, introduces a more general framework for defining discourse based on rationality. He posits that a conversation becomes a discourse when a rational motivated consensus could be achieved through the presentation of arguments without restrictions on how often or how long these arguments are presented (Habermas, 1981a: 71). This concept aligns well with conversations on social media platforms, which do not impose such limitations. In contrast to live debates, discussions on social media are not subject to the physical constraints of human fatigue and can stretch indefinitely, assuming the platform remains available, and attention remains engaged.

This definition of deliberation introduces consensus-seeking behavior as a requirement for deliberation. It also underscores the role of individual participants, emphasizing the need to transcend egocentrism to achieve a discursive consensus. But even the vaguest progress towards consensus can only be made if views are exchanged which implies a certain length of the discussion. We will build on this assumption without providing evidence as evaluating consensus seeking behavior empirically is challenging at least (Steenbergen 2003; Bächtiger & Parkinson, 2019: 50).

After defining deliberation as a directed process towards a goal, we investigate whether conversational progress is possible in the structures of social media. As an analogy with offline-conversations, we are interested in longer reply-chains as depicted in figure 1. Here, the nodes are the posts, and the edges read from top to bottom as a post answering another post. The root of the tree is the original post in the online conversation. Every online forum and social media thread can be modeled this way because every post except the root post has a parent which is the mathematical definition of a recursive tree structure.

Figure 1 (left): A discussion tree in social media with the nodes as posts, the edges denoting replying, and the root of the tree as the original post. Figure 2 (right): The Mushroom Shape of typical Twitter Conversation Trees.



The marked path is one of many pathways that can be written down like a transcript from a group discussion. Pathways can be defined as all the paths in a tree that start with the root and end in a leaf (a node without children). This approach serves the function of filtering linear reply-chains in social media (Wang 2008; Nishi 2016), that can be considered an online equivalent of real-life discussions.

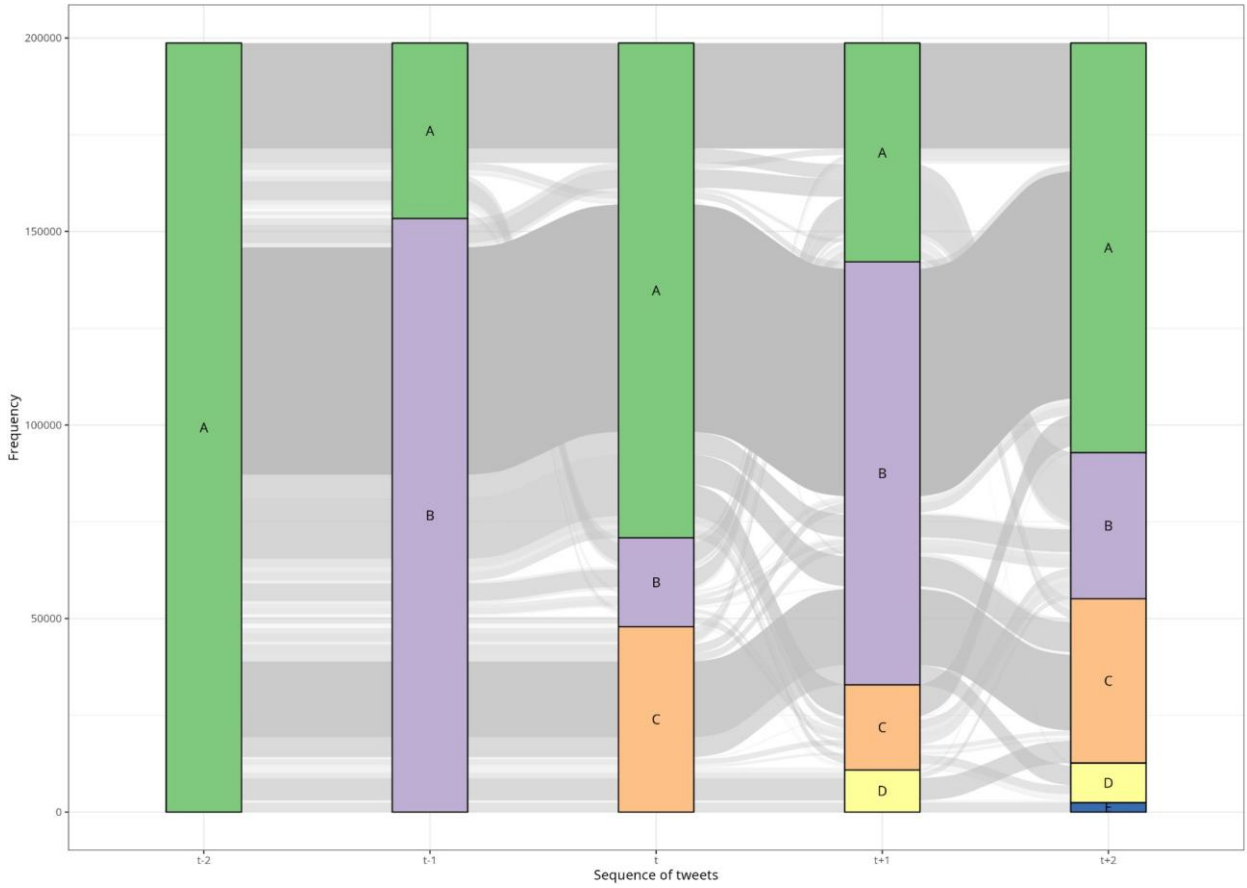
The question is one of definition: we posit that deliberation cannot take place if one post is answered one time only but needs a longer exchange of arguments and perspectives. This back and forth is needed to shed light on the subject and to allow consensus to form or at least to explain where everybody stands. Given our interest in moderation we define a minimal length of five consecutive posts

for the following reasons: (a) at least two posts are needed to initiate a dialogue, (b) a third post is needed that contains the moderation attempt according to the process function discussed earlier, (c) two more posts are needed to give the first two authors a chance to react each. Thus, the minimal five-post window is artificial but not arbitrary: it follows logically from the requirements of deliberative talk of reciprocity, exchange of arguments and of it being a group activity. It also takes into account the limitations of attention span in social media. It can be safely assumed that very few users read all previous posts in a discussion thread. Even less follow all the branches of the discussion to the end. This means that the five-post window may not only be the minimal length but also the maximum. Preliminary research suggest an attention span of three posts (N.,N. 2032, ArchivX link). However, studying the exact boundaries of user attention is not possible in this context, so the question of how to segment longer discussions has to be omitted. As a last argument from a research-pragmatic point of view it can be argued that a sequence of 5-posts also proved to be a reasonable chunk of information for coders to process at a time.

Another question is the number of participants involved. Anticipating the sample of the case study described later in this paper, Figure 3 shows the patterns of author identities within a 5-post window.

This overview over the patterns shows that even the longer reply chains are mostly a result of two author's talking to each other (a dialog) rather than a group discussion assuming that a social group consists of more than two. It also means that in the majority of the cases the author of the moderating post has written one of the two previous posts. As the moderator is part of the dialogue, the definition of partial moderation is explained structurally: moderation from ordinary users is mostly a byproduct.

Figure 3: Every vertical line represents the position in the 5-window.



In summary, the idea of social media discussions as deliberation is misleading. Broadcasting, short responses and dialogues do not fulfill the requirements set by the normative theory. The reply-trees need to be filtered for longer pathways and these need to be filtered for higher numbers of distinct authors. Only then the decidedly fewer instances of discussions can be investigated for the deliberative democracy research program.

6. Case Study of User Moderation Strategies on Twitter

The following gives an overview over the research process and the interaction between the intermediary results. The case study consisted of four phases: 1. Deduction, 2. Sampling, 3. Resampling and 4. Qualitative Analysis. In the first phase the developed definition of moderation as a *process* was used to deduce identifiable moderation strategies that support deliberation online. During the deduction phase the intervention examples were generated based on the ideal of deliberation (consensus-seeking, rationality, etc.), the conversational norms that are enforced by content moderation (politeness, civility, openness, etc.) and the structure of internet conversations as detailed in section five. These candidates were also classified using the Moderation Context Model (Figure 1). With this approach we arrived at 13 gold candidates and 40 lesser candidates (Figure 4) for deliberative moderation strategies.

In the second phase, the strategies were used as a template to find search phrases for social media. Using an abductive process that involved checking for presence in social media, arguing about whether the phrase represents the process function of the moderation and checking for linguistic permutation of the syntax, corresponding search phrases were generated. These phrases were used to download online conversations, to label the resulting posts according to our definition of what moderation is and to investigate which strategies can be identified with a high intercoder reliability. This process was repeated until a pattern emerged that pointed to which categories of moderation strategies could reliably be found.

In the third phase the reliable phrases were sorted into three main categories, and used to download a large corpus consisting of full conversations. Using the concept of the conversation pathways (Figure 1) and computational pathfinding, linear conversation transcripts were extracted from the conversation trees. This resulted in 490 discussion transcripts. The corpus was also used to analyze the patterns of

user responses which resulted in the observation of the prominence of dialogues as discussed previously.

In the fourth and final phase these discussion transcripts were investigated qualitatively whether the moderation phrase signified the intent to moderate the discussion as defined by the partial moderation concept. Afterwards, the issue-relationship dichotomy was labeled in order to assess whether the post had the effect of steering the conversation into more rational waters. This contrasts the literal interpretation with the pragmatic one which is based on the effect of a speech act. Finally, the categories from phase one and three were cross-checked for their validity and exclusiveness. In the following these phases will be elaborated in more detail.

6.1 Deduction

In order to arrive at a framework for deliberative moderation that works with social media, structured deductive content analysis (Meyring 2004) was employed. Instead of starting with a greenfield approach, we deduced phrases of moderation that could be used for sampling user moderation using the Twitter API. In the following it will be described how these phrases were generated based on the moderation theory, the deliberation theory and the different meanings of moderation depending on the social context. Without knowing the culture of social media moderation (unitary or adversarial, issue focused, or relationship focused) the starting point was the question whether you should assume an unitary or an adversarial concept of man (*Menschenbild*).

	Unitary Frame	Adversarial Frame
High Relationship & Low Issue Focus	Improve relationship, Understand/appreciate another's perspective, …	Support elaboration, Tone Policing, End the discussion peacefully, …
Low Relationship & High Issue Focus	Agenda Control, Support Consensus Building, invite new perspectives, Summarize important points, …	Point out commonalities, Point of differing interest structures, Increase Bargaining Space, Find a compromise …

Figure 3: List of Moderation Strategies based on Construction

Whether the disposition to seek common ground is dominant or interests play a major role cannot be tested easily in social media, as participants are anonymous. In this study the number of conversations exceeds the limit of what is feasible with questionnaires. But in each hypothetical case deliberative moderation is useful: if social media interactions were purely adversarial, the role of the moderator in that case would be to function as a bridge or facilitator in order to steer the process towards consensus seeking even if this ideal goal is achieved through minimal bargains. If consensus seeking was part of social media, the role of the moderator would be similar to the chair in an academic panel trying to improve the result by summarizing, agenda control or such. If common ground already exists and the process is not open for participation because of the phenomenon of echo chambers (Flaxman & Goel & Rao 2016) the moderator would have to foster the reflection of belief systems by engaging in Socratic dialogue which means to ask questions about the assumptions and shake the belief of knowing a one-sided truth.

With this line of reasoning, we apply the ideas of the process function and the deliberative goals of consensus-seeking in order to generate useful strategies of moderation. Examples of these are listed in

Figure 3. With the Y-Axis, the dichotomy of issue-focus and relationship-focus, there is no need to hypothesize as this can be observed directly in the discussions in social media.

6.2 Sampling

The study uses Twitter data between Februar 2022 and February 2023. English and German Tweets were queried using the Twitter API. The language can sometimes be classified as a mix of English and German.

The generated strategies were operationalized as search phrases that could be used to download conversations from social media and build the corpus. Figure 4 illustrates how the strategy was converted into searchable phrases. Herein lies the main limitation of the study in that even with the support of automated phrase generation not all possible linguistic expressions of the strategies are generated. This limits the generalizability of the created corpus. However, this process can still be viewed as an improvement over other approaches for the following reasons:

1. No assumption was made of what constitutes deliberative quality
2. There is no need for proxies such as sentiment for tone, or argument density for rationality
3. The role of the users is not coded implicitly into the corpus, for instance by sampling moderation by similarity to speech acts by moderators in TV shows or users with an explicit moderator role which exists in the platform Reddit.
4. The phrases are linguistic features and can be searched on any platform.
5. The phrases can also be identified for short exchanges. Other automated measures need a bigger text input to be reliable.

Furthermore, the phrases are added in the appendix and can be extended by future researchers in order to reproduce the results or to study other aspects of deliberative user moderation.

Moderating Strategy	Example Search Phrase
Pointing Out Commonalities	You are working towards a similar goal; You come from a similar background
Find a Compromise/End Discussion Peacefully	Let's agree to disagree
Understand/Put Oneself in Another's Shoes	You have a valid point
Prompt to Elaborate on One's Opinion	Can you maybe explain that further; Why do you feel that way?
Improve Relationship	Why are you so personal about this? All here are human and have feelings (except me, I am a bot); It does not make sense to make this personal with everyone being anonymous!
Refocus on the Topic	Perhaps we could get back to the main point; Let's focus on the Issue at hand; Stay on the topic, please! What are you talking about! Can you explain it to me?
Tone Policing	Let's focus on Constructive Criticism! Insulting each other does not help anyone.
Point of Differing Interest Structures	What do you hope to get out of this? What's your intention/reasoning?
Increase Bargaining Space	Are there more aspects that need to be included; In a perfect world, what would be the biggest common denominator? How do we move forwards?
Support Consensus Building	What are common assumptions you share? Which of the arguments presented above is better? What can we agree upon?
Summarize Important Points	My takeaways from this are; To recap what has been said; in summary…
Invite New Perspectives	I am happy to hear any opinion; curious to hear; any comments; any questions
Tone Policing (Civility focus)	Let's keep the discussion civil

Figure 4: Examples of Search Phrases based on Moderation Strategies

To arrive at a sufficiently large dataset with conversations containing moderation instances, two challenges remain: first, not all phrases clearly indicate moderation; there is some ambiguity involved. Hence, we chose to label moderation instances based on a judgment made by two annotators. The 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.

Table 1: Most frequent strategies with highest intercoder reliability (elaboration support, emotion control and agenda control)

query category	**n_Units**	**n_Coders**	**Holstis_CR**	**Krippendorff's Alpha**	**Fleiss Kappa**
elaborate *	124	2	0,903	0,717	0,716
topic	155	2	0,852	0,689	0,688
explain *	224	2	0,844	0,623	0,622
valid_point	79	2	0,899	0,545	0,542
agree	88	2	0,909	0,508	0,505
* queried in English and German				** insufficient data	

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 0.95 (Appendix 2). The inter-annotator agreement is based on Krippendorff's alpha (Krippendorff 2004). The disagreement is mainly due to difficulties in interpreting the degree of metatalk. As this is only an intermediary step, the inter-annotator agreement does not need to be high enough for inference but serves as an indicator that the phrase works as a proxy to find moderating instances reliably.

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 label 1.500 tweets manually. We then query those phrases at a large scale that show both a high inter-annotator agreement (greater than 0.6) as well as a high precision (greater than 0.1) for moderation instances showing up at the right position, as we define a moderating post to follow at least two posts and be followed by two posts (see discussion on minimal length of pathways earlier). First, there needs to be something to moderate and, second, we need a minimal reaction in order to assess the effect of the post as moderating.

6.3 Resampling and Answer Pattern Analysis

Utilizing a selected group of 7 phrases, we collected approximately 140,000 tweets in English (127,782) and German (11,776), each containing at least one of these phrases. These tweets were posted from February 2022 to February 2023, and we anticipate some variations over this period. There is a noticeable imbalance in the distribution of these phrases – both across different moderation categories and languages. Notably, around 110,000 English tweets predominantly fall under the category of engaging participation. In contrast, the most frequent phrase in German tweets is associated with tone policing.

From this collection, we extracted a random subset to obtain the complete conversations, resulting in a dataset comprising 39,371 unique dialogues totaling 4,298,811 tweets. Ideally, these conversations would represent the complete dataset for analysis in a perfect social media environment. However, not all participants respond directly to the previous tweet; some reference the discussion through hashtags or @-mentions, leading to approximately 10,000 conversations that contain only a single tweet. Additionally, there are instances where earlier tweets in a thread have been deleted, causing some conversations to begin with the second or third tweet. To address these irregularities, we refined our

dataset to include only those conversations with more than 5 and fewer than 500 tweets. Consequently, our dataset was reduced to 25,070 unique conversations, encompassing 4,126,948 tweets.

6.4 Qualitative Content Analysis

Despite the high number of tweets containing the search phrases in the sample, only few cases remained that fulfilled all the requirements explained in section 5. After filtering for exchanges that lasted at least five posts that include the moderation in the middle and using only those phrases that had a high intercoder-reliability, only 490 cases remained. These were hand-coded according to the categories developed from the theory.

Table 2: Result of the labeling process in terms of intercoder-reliability.

Moderation Strategy	**Count**	**Issue-Focussed Reaction in %**	**Cohen's Kappa**
Elaboration Support	342	.92	.95
Emotion Control	68	.63	.72
Agenda Control	79	.72	.84

Table 2 above shows the distribution of the strategies in that final sample. Because the sample already contained the maximum of the data points available, little could be done about the skewed distribution. This also favored the qualitative approach. Appendix 4 shows the coding scheme for the initial categorical coding. The main purpose of this was to validate the sample as instances of user moderation by assessing whether they had the effect on the subsequent discussion of being more issue-focussed. It turns out that the moderation had this effect on the conversation in the majority of cases. The previous finding that the search phrases signified the intent of moderation was also confirmed. In the subsequent case analysis, the constructions of *partial moderation* (section 3), the process character (section 4) as well as the assumed position in the MCM were investigated and validated.

In the following the three distinct moderation strategies are illustrated with one in depth example each. Because of legal reasons the examples have been rewritten manually and using Large Language Model-Paraphrasing to make it impossible to find the original conversation on Twitter:

Example for Agenda Control:

> ***Primary statement:*** *I believe individuals should value others' perspectives. Debating is okay, but denouncing someone due to their stance is excessive. Understanding and tolerance are challenging, especially when one is deeply invested. Nonetheless, one's stance is their own. See the link: [URL here]*
>
> *…*
>
> ***1. @userA:*** *Saying "every view has merit" is oversimplified nowadays. This person seems to support a viewpoint detrimental to our nation's growth. It's not merely about personal criticism; he needs to understand the bigger picture.*
>
> ***2. @userB:*** *I don't see eye to eye with this. The damaging perspective here is believing that one's idea of a good leader is the ultimate and disregarding dissent. Isn't choice what democracy stands for? I'm not in favor of suppressing diverse viewpoints.*
>
> ***3. @userA:*** *Stick to the topic. He's not being criticized for endorsing the XYZ leader. Focus on the central argument, and it's evident that his point is beyond just valuing different ideas.*
>
> ***4. @userA:*** *He's echoing a sentiment that's degraded our nation. Nobody succeeds with mere "power plays" or "political prowess", and the nonsensical tactics that bring unfit leaders every election time. Such narratives need to be challenged.*
>
> ***5. @userB:*** *Managing elections demands resources; it's quite straightforward. In certain situations, the public willingly contributes. That's just how things work. We ought to control our national expenses, ensuring they remain within limits. Globally, electoral costs have to be managed.*

This is the format the coders encountered the cases. The first post is the root post defining the conversation. After that there may or may not be a break until the point where the five-post window starts, where the moderating phrase is in the middle. In this case it is 'stick to the topic'. According to the self-imposed rules (codebook in Table A.2) there is *metatalk* that tries to take on a bird-eye view of the discussion, arguing that the last comment was off topic without breaking social norms. The same

comment could be viewed as a lot less productive if it was accompanied by insults or if the criticism was not explained that well. Regarding the question of *partial moderation*, this example illustrates that the moderating sentence is not disjunct from the general point the author is making. Still the message falls exactly into the intent of agenda control and can be defined as such.

This example also illustrates the gray line of interpreting the user moderation *with the intention* of improving the conversation instead of judging its moderating quality *by the effect*. The third dimension of the interpretation is concerned with the question how literal the text is to be understood. Most of the time, there is a clear rhetorical purpose like agenda setting, gaining moral high ground or employing rhetorical devices. The pragmatic interpretation focusing the intent was prioritized here. Although the goal of staying with the topic might not have been fulfilled the intent was to steer into that In this case one might conclude that the post was not moderating but the opposing statement ' He's not being criticized for endorsing XYZ leader.' was bolstered by the academic norm of staying on topic. However, here the statement 'focus on the central argument' underlines the message and makes this a clear case for the coders. Finally, this case has a clear issue-focus. So the moderation takes place in the area of civil talk and tries to optimize the process. This fits with our classification of agenda control in the bottom left quadrant of the MCM.

Example for Emotion Control:

> ***@RootPoster****: Just watched a video where someone was mistreating an animal while others just laughed. How can people differentiate between the value of a human and an animal's life? Aren't both important? 😔 #AnimalRights #Humanity*
>
> *…*
>
> ***@UserA****: This disgusting kid, how can someone be so stupid? I get so aggressive when I see this.*
>
> ***@UserB****: Dude, what's wrong with your head? I've never seen such pure stupidity, oh my god.*
>
> ***@UserC****: Calm down, B.. He already wrote what he meant.*
>
> ***@UserD****: I'm sorry, but in a league where rapists and racists play, this statement is catastrophically stupid. Of course, one should never hit an animal, but at the end of the day, it's an animal, dude, not a human.*
>
> ***@UserE****: The animal can defend itself even less. The animal can't do anything; both are bad and should be condemned.*

In this example the moderation strategy was to focus on the issues and foster common understanding. In this case UserC does not position herself in the argument so that it can be safely assumed that there is a genuine interest in calming down the discussion. In this case the effect is immediate and subsequent posts even though emotionally heated present arguments instead of attacking the speaker. When investigating the question of *partial moderation* the main content of the post addresses the question of the tone or summarizes previous events. In this case the post could be labeled fully moderating. In terms of classification with the MCM the assumption here is that there is a relationship focus and adversarial dispositions. This also fits our assumption placing emotion control in the top right corner, trying to steer the conversation in a less adversarial and less relationship focused direction (toward bottom left). As the post also has an organizing function (“He already wrote what he meant”) the requirement of the definition of moderation as “structuring the process” is fulfilled, too. This is crucial to differentiate this from counter-speech.

Example for Elaboration Support:

> ***@RootPoster**: Just saw a discussion on why restaurant visits declined again. Was it the omicron variant or the fucking government policies like the mask mandates that caused it?*
>
> ***@UserA**: What?*
>
> ***@UserB**: Where?*
>
> ***@UserC**: Can you elaborate on how it was the gov that made people not want to go rather than, say, idk, the virus spiking?*
>
> ***@RootPoster**: Bc they made masks political. Once they brought the mask mandate back, the servers suffered because people stopped going out. This combined with the unemployment taken away caused millions of servers and bartenders to suffer. Does that help?*
>
> ***@UserE**: The mask mandate didn't stop people from going to restaurants *where they didn't have to wear masks. The industry suffered due to omicron. Had there been no omicron, the industry would have bounced back a lot.*

In this case the original post was a bit more polarizing than the smoothed version we are presenting here. The moderating intent of UserC here is to bring out the hidden assumptions and political stance of the original post which in this case came directly before the five-post window. Without any non-verbal clues and anonymity these requests to express something more in detail can have the effect of inviting participation to get to the bottom of things or to give a speaker a chance to explain their position.

It also is a way of introducing a new perspective in an ingroup setting where the first three participants probably share a certain view on Covid policy whereas the UserC and UserE represent another stance.

The example has a focus on information gain or fact checking. The effect of questioning assumptions is only secondary. Here, the response shows that explanations and exchange of views follow which makes it easier to classify this as moderating. The example also shows why a qualitative approach makes sense: as the unintended effect, the literal intention and the possible deeper interpretation have to be weighed against each other, coding this quantitatively would run the risk of forcing the biased result which is finding ordinary users moderating.

In summary, the MCM framework successfully captured the cases and the phrase based sampling approach could be validated. The case study also showed the limits of the concept acknowledging that even the very few hand-picked cases were far from the ideal deliberation. Logically, user moderation was always subordinated to the personal agenda.

7. A Framework for Deliberative Moderation

Going through the different combinations of process functions of moderation, their connection to deliberation, their codable appearance in social media and their place in the MCM, five moderation strategies for deliberative moderation in social media were distilled.

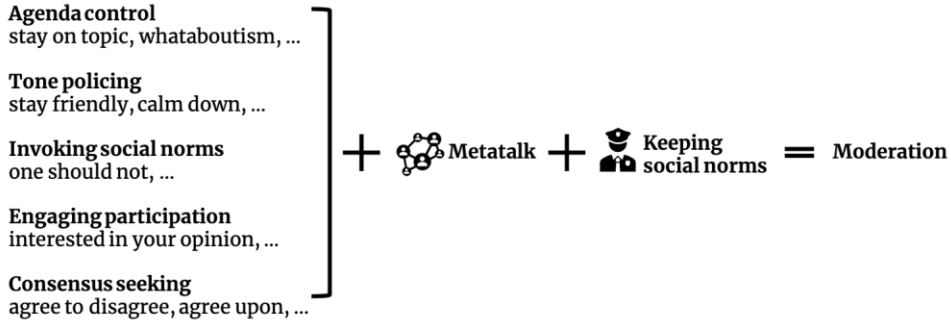


Fig 3: Summary of Moderation Strategies

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. Both stress the importance of ensuring a sensible process for arguments to come out.

8. Discussion and Concluding Remarks

There are some limits to applying the MCM to social media. It is unclear what to expect regarding the level of cooperation within the community. A further unanswered question is, how closely the community is linked and how clear-cut are the fringes of the community. In short, the adversarial disposition could not be generalized. Both ends of the spectrum exist.

The vast majority (over 99%) of reliably coded strategies are one of agenda control, tone policing, and engaging participation. There are two explanations for this finding: it is possible that the

motivation, and the structure of Twitter conversations is not suited to develop an overarching view of the discussion. Another plausible argument is that this kind of behavior is more complex and thus harder to agree upon during the labeling process. Especially the category *norm control* showed a low intercoder reliability as there was a disagreement between the coders whether a moderating statement needs to be tone-neutral or not. Lastly, this might be due to the phrases working better as a filter for the simpler strategies.

From the coding, it is not clear whether elaboration support and engendering participation should be two distinct categories. Whilst social media are blind to the composition of different actors, genders or ethnic groups this might still be an issue if the model was transferred from social media to offline deliberation. Engendering participation could be inferred from the author patterns discussed: dialogues could be considered less participative than a third or fourth person joining. Methodologically, this would mix the qualitative approach with a quantitative measure. For this reason, it was omitted here but would be an interesting next step in the analysis.

Often moderation is not due to the intrinsic motivation of improving the conversation but used as a rhetorical device or power play. It happens a lot to find a moderative comment that structures the conversation nicely but is highly intolerant at the same time. This result helps to investigate ethical gray areas of the deliberation concept such as the conflict between the process function and norms of civility. Twitter, Facebook, and other similar platforms claim to be spaces that support free speech and democratic exchange. However, the democratic public sphere demands more than just the freedom to speak. It requires a space where participants engage in reasoned debate of a minimal length and constructive dialogue.

The main contributions of the paper are theoretical: (RQ1) *deliberative moderation* has been conceptualized as an intervention that tries to moderate the process of deliberation. *Partial deliberative moderation* has been defined in order to make the definition useful for empirical analysis. The

Moderation Context Model (MCM) has been drafted as both a theoretical contribution to deliberation research as well as an instrument for empirical analysis which allows to place different strategies of deliberative moderation based on their intent and context. The MCM provides a promising foundation to create deliberative moderation schematics for various application fields. If the current taxonomy of moderation strategies proves to be too narrow, this framework offers a starting point for adaptation and generalization.

(RQ2) The interpretation of deliberation as a process was applied to social media and longer reply threads in social media were analyzed based on their structure and whether they fulfill the requirement to be considered a deliberative discussion. As an unintended side effect the nature of Twitter conversations as mainly not-deliberative was discovered on two counts: first, the presence of longer reply-chains was found to be rare; secondly, the longer reply-chains represented dialogues mostly, which puts the assumption to the test that deliberation, or a public sphere, can be conceived as a group discussion in social media. This study also provided methodological insights, in how to process and filter social media data, to find instances that can sensibly be analyzed under the heading of deliberative research.

(RQ3) In the case study, the theoretical concepts were both validated and elaborated upon. From this a taxonomy of deliberative moderation strategies was derived. An important takeaway from the case study is that deliberative moderation is not an intervention that suddenly improves deliberative quality but it is an attempt to steer certain types of discussions towards an ideal type of discussion. This happens organically on social media when users engage more deeply.

Although the existance of deliberative moderation was shown on social media, its rare nature should not be ignored. Other platforms with differing structures and greater average lengths of discussions should be investigated to answer the question if online deliberation needs professional moderators or if the market of ideas can regulate itself.

9. Data Accessibility

According to the legal terms of the Twitter API, the conversations and their codes are not published. Examples are obfuscated and cannot be traced back to the authors.

However, the software developed to download and analyze the data before the qualitative coding is available as open source on Github, and can be used under the MIT license: [anonymized link] contains the code to download the conversational data. [anonymized link] contains the code for analyzing the reply-trees and identifying longer reply-chains that are used as discussions. [anonymized link] contains the repository for the preliminary study that tried to identify user moderation automatically.

10. Funding Information

To be completed after review

11. Authors' contributions

To be completed after review

12. Competing Interests

To be completed after review

13. References

Adler, Richard P., und Judy Goggin. 2005. „What Do We Mean By "Civic Engagement"?" *Journal of Transformative Education* 3 (3): 236–53. https://doi.org/10.1177/1541344605276792.

Alsagheer, Dana, Hadi Mansourifar, und Weidong (Larry) Shi. 2023. „Statistical Analysis of Counter-Hate Speech on Voice-based Social Media". *Procedia Computer Science*, The 14th International Conference on Ambient Systems, Networks and Technologies Networks (ANT) and The 6th International Conference on Emerging Data and Industry 4.0 (EDI40), 220 (Januar): 1009–14. https://doi.org/10.1016/j.procs.2023.03.140.

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

Bandler, Richard, and Robert Dilts. 2015. "Reframing." In Grundlagen der Kommunikation. Utb.
Bächtiger, André, and John Parkinson. 2019. Mapping and Measuring Deliberation. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780199672196.003.0002.

Bonaldi, Helena, Sara Dellantonio, Serra Sinem Tekiroglu, and Marco Guerini. 2022. „Human-Machine Collaboration Approaches to Build a Dialogue Dataset for Hate Speech Countering". arXiv. https://doi.org/10.48550/arXiv.2211.03433.

Bormann, Marike, Dominique Heinbach, and Marc 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, 5. DEU.

Coe, Kevin, Kate Kenski, und Stephen A Rains. 2014. „Online and Uncivil? Patterns and Determinants of Incivility in Newspaper Website Comments". Journal of Communication.

Chambers, Simone. 2003. "Deliberative Democratic Theory." Annual Review of Political Science 6(1):307-326. https://doi.org/10.1146/annurev.polisci.6.121901.085538.

Darmody, Merike, and Kerzi, Jenniferl. 2016. Education Policies and Practices to Foster Tolerance, Respect for Diversity and Civic Responsibility in Children and Young People in the EU: Examining the Evidence. LU: Publications Office of the European Union. https://data.europa.eu/doi/10.2766/46172

Doğanç, Mekselina, and Ilia Markov. 2023. „From Generic to Personalized: Investigating Strategies for Generating Targeted Counter Narratives against Hate Speech". In Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA), 1–12. https://aclanthology.org/2023.cs4oa-1.1/.

Duong, Phuc H., Cuong C. Chung, Loc T. Vo, Hien T. Nguyen, and Dat Ngo. 2021. „Detecting Hate Speech Contents Using Embedding Models". In Computational Data and Social Networks, herausgegeben von David Mohaisen und Ruoming Jin, 138–46. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-91434-9_13.

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.

Epstein, Dmitry and Leshed, Gilly (2016) "The Magic Sauce: Practices of Facilitation in Online Policy Deliberation," Journal of Public Deliberation: Vol. 12: Issue 1, Article 4.

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, 7. DEU.

Friess, Dennis, Marc Ziegele, andDominique Heinbach. 2021. „Collective Civic Moderation for Deliberation? Exploring the Links between Citizens’ Organized Engagement in Comment Sections and the Deliberative Quality of Online Discussions“. Political Communication 38 (5): 624–46. https://doi.org/10.1080/10584609.2020.1830322.

Fisher, Roger, W. U., and B. Patton. 2015. “Kommunikation und Verhandeln.” In Grundlagen der Kommunikation: Gespräche effektiv gestalten. Utb.

Flaxman, Seth, Sharad Goel, and Justin M. Rao. 2016. “Filter Bubbles, Echo Chambers, and Online News Consumption.” Public Opinion Quarterly 80: 298-320.

Ganesh, Bharath, and Jonathan Bright. 2020. „Countering Extremists on Social Media: Challenges for Strategic Communication and Content Moderation“. Policy & Internet 12 (1): 6–19. https://doi.org/10.1002/poi3.236.

Garimella, Kiran, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. „Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship“. In Proceedings of the 2018 World Wide Web Conference, 913–22. WWW ’18. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3178876.3186139.

Garrett, R. Kelly. 2009. “Echo Chambers Online? Politically Motivated Selective Exposure among Internet News Users.” Journal of Computer-Mediated Communication 14(2):265-285. https://doi.org/10.1111/j.1083-6101.2009.01440.x. Accessed February 7, 2022.

Graham, Todd, und Tamara Witschge. 2003. „In search of online deliberation: Towards a new method for examining the quality of online discussions“. *Communications* 28: 173–204.

Grube, Joel W., Daniel M. Mayton II, und Sandra J. Ball-Rokeach. 1994. „Inducing Change in Values, Attitudes, and Behaviors: Belief System Theory and the Method of Value Self-Confrontation“. Journal of Social Issues 50 (4): 153–73. https://doi.org/10.1111/j.1540-4560.1994.tb01202.x

Guerra, Pedro, et al. 2013. “A Measure of Polarization on Social Media Networks Based on Community Boundaries.” Proceedings of the International AAAI Conference on Web and Social Media 7(1):215-224. Accessed February 10, 2022.

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

Habermas, Jürgen. 1981a. Theorie des Kommunikativen Handelns: Band 1. Frankfurt am Main: Suhrkamp.

Habermas, Jürgen. 1981b. Theorie des Kommunikativen Handelns: Band 2. Frankfurt am Main: Suhrkamp.

Habermas, Jürgen. 1982. Strukturwandel der Öffentlichkeit. Frankfurt am Main: Suhrkamp.

Habermas, Jürgen. 2009. Diskursethik. Frankfurt am Main: Suhrkamp.

Karpowitz, C., and Mansbridge, J. 2005. Disagreement and Consensus: The Importance of Dynamic Updating in Public Deliberation. In Gastil, J. and Levine, P. The Deliberative Democracy Handbook: Strategies for Effective Civic Engagement in the Twenty-First Century. San Francisco. Jossey-Bass.

Kim, Nuri. 2016. „Beyond Rationality: The Role of Anger and Information in Deliberation". Communication Research 43 (1): 3–24. https://doi.org/10.1177/0093650213510943.

Krippendorff, K. 2004. "Reliability in Content Analysis." Human Communication Research 30 (3): 411-433.

Kuckartz, U. 2016. Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung. 3rd ed. Weinheim: Beltz Juventa.

Löb, Charlotte, und Hartmut Wessler. 2021. „Mediated Deliberation in Deep Conflicts: How Might Deliberative Media Content Contribute to Social Integration Across Deep Divides?" Journal of Deliberative Democracy 1 (1). https://doi.org/10.16997/jdd.981

Mansbridge, J. 1980. Beyond Adversarial Democracy. Chicago. University of Chicago Press.
Mayring, P. 2014. Qualitative content analysis: theoretical foundation, basic procedures and software solution. Klagenfurt. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-395173.

Mayring, P. 2015. Qualitative Inhaltsanalyse: Grundlagen und Techniken. 12th ed. Weinheim; Basel: Beltz.

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

Monnoyer-Smith, Laurence, und Stéphanie Wojcik. 2012. „Technology and the Quality of Public Deliberation: A Comparison between on and Offline Participation". *International Journal of Electronic Governance* 5 (1): 24–48.

Muhlberger, Peter. 2005. "The Virtual Agora Project: A Research Design for Studying Democratic Deliberation." *Journal of Public Deliberation* 1 (1): Article 5.

Mutz, Diana C. 2008. „Is Deliberative Democracy a Falsifiable Theory?" *Annual Review of Political Science* 11: 521–38. https://doi.org/10.1146/annurev.polisci.11.081306.070308.

Nishi, R., T. Takaguchi, K. Oka, T. Maehara, M. Toyoda, K.-i. Kawarabayashi, and N. Masuda. 2016. “Reply Trees in Twitter: Data Analysis and Branching Process Models.” Social Network Analysis and Mining 6 (1): 26.

Papacharissi, Zizi. 2004. „Democracy Online: Civility, Politeness, and the Democratic Potential of Online Political Discussion Groups“. New Media & Society 6 (2): 259–83. https://doi.org/10.1177/1461444804041444.

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.” (preprint).

Rapp, Carolin, und Markus Freitag. 2015. „Teaching Tolerance? Associational Diversity and Tolerance Formation“. Political Studies 63 (5): 1031–51. https://doi.org/10.1111/1467-9248.12142.

Rombout, F., J. A. Schuitema, und M. L. L. Volman. 2022. „Teaching Strategies for Value-Loaded Critical Thinking in Philosophy Classroom Dialogues“. Thinking Skills and Creativity 43 (März): 100991. https://doi.org/10.1016/j.tsc.2021.100991.

Rosenberg, Marshall B. 2016. Gewaltfreie Kommunikation: Eine Sprache des Lebens. Junfermann Verlag GmbH.

Rowe, Ian. o. J. „Deliberation 2.0: Comparing the Deliberative Quality of Online News User Comments Across Platforms“. *Journal of Broadcasting & Electronic Media* 59 (4): 539–55. https://doi.org/DOI: 10.1080/08838151.2015.1093482.

Ruin, S. 2017. “Ansätze und Verfahren der Kategorienbildung in der qualitativen Inhaltsanalyse.” In Schulsportforschung–wissenschaftstheoretische und methodologische Reflexionen, 119-134.

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.

Spada, Paolo, and James Raymond Vreeland. 2013. “Who Moderates the Moderators? The Effect of Non-Neutral Moderators in Deliberative Decision Making.” *Journal of Deliberative Democracy* 9 (2). https://doi.org/10.16997/jdd.165.

Steenbergen, Marco R., et al. 2003. “Measuring Political Deliberation: A Discourse Quality Index.” Comparative European Politics 1(1):21-48.

Stromer-Galley, Jennifer. 2007. “Measuring Deliberation’s Content: A Coding Scheme.” Journal of Public Deliberation 3(1): Article 12.

Stroud, Natalie Jomini, Joshua M. Scacco, Ashley Muddiman, und Alexander L. Curry. 2015. „Changing Deliberative Norms on News Organizations' Facebook Sites". Journal of Computer-Mediated Communication 20 (2): 188–203. https://doi.org/10.1111/jcc4.12104.
Tekiroglu, Serra Sinem, Yi-Ling Chung, und Marco Guerini. 2020. „Generating Counter Narratives against Online Hate Speech: Data and Strategies". arXiv. http://arxiv.org/abs/2004.04216.

Tucker, Joshua, et al. 2018 "Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3144139.

Türkkahraman, Mimar. 2014. „Social Values and Value Education". Procedia - Social and Behavioral Sciences, 5th World Conference on Educational Sciences, 116 (Februar): 633–38. https://doi.org/10.1016/j.sbspro.2014.01.270.

Ury, William. 1993. "Getting Past No: Negotiating in Difficult Situations". Bantam.

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.

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, 1–13. New York, NY: Association for Computing Machinery.

Yildirim, Mustafa Mikdat, Jonathan Nagler, Richard Bonneau, und Joshua A. Tucker. 2021. „Short of Suspension: How Suspension Warnings Can Reduce Hate Speech on Twitter". Perspectives on Politics, 1–13.

Ytre-Arne, Brita, and H. Moe. 2018. "Approximately Informed, Occasionally Monitorial? Reconsidering Normative Citizen Ideals." The International Journal of Press/Politics 23:227-246.

Ziegele, Marc, und Pablo B. Jost. 2020. „Not Funny? The Effects of Factual Versus Sarcastic Journalistic Responses to Uncivil User Comments". Communication Research 47 (6): 891–920. https://doi.org/10.1177/0093650216671854.

Ziegele, Marc, Teresa K Naab, und Pablo Jost. 2020. „Lonely Together? Identifying the Determinants of Collective Corrective Action against Uncivil Comments". New Media & Society 22 (5): 731–51. https://doi.org/10.1177/1461444819870130.

Zuiderveen Borgesius, Frederik J., et al. 2016. "Should We Worry About Filter Bubbles?" Internet Policy Review 5(1). https://doi.org/10.14763/2016.1.401.

Appendix 1: Rules for Moderation Classification

Rules for coding moderating instances:

A: Metatalk

- "This conversation is going nowhere. We should focus on these XY aspects."
- "Everyone here is always taking this side X."

B: Structuring the conversation

- "Let me summarize the last tweets, so we can come to a conclusion."
- "There seem to be these three positions."

C: Invoking social norms

- "I don't agree with your sentiment concerning Sinti and Roma."
- "You should not insult people if you want to lead a decent conversation."

D: Engaging participation

- "Can you elaborate further?"
- "What do you (author of tweet 3x before) think about it?"

E: Not breaking social norms

F: Breaking social norms in an extreme way

- Heavily insulting minority groups, e.g.: "you damn Jews"; "you damn Gypsies."

G: Consensus-seeking

- "Can we agree that the rally was fraud?"
- "Can we agree not to conduct party-political disputes at the expense of the credibility of public-law broadcasting?"
- "We can agree on that."

H: Shifting emotional mode to rational/issue-based forms

- "Calm down."
- "Relax. Stay on the issue at hand!"

Rules:

1. If A (and not F), then M.
2. If B (and not F), then M.
3. If C and somewhat A or B, then M.
4. If D and E (and not F), and somewhat A-C, then M.
5. If G and E (and not F), and somewhat A-C, then M.
6. If H and E, then M.

Appendix 2: Full Table of Coded Phrases

moderation_type	query	n_Units	Agreement	Holstis_CR	Krippendorffs_Alpha	Fleiss_Kappa	freq_positive_cases	freq_negative_cases
agenda control_de	Thema bleiben	12	0,833	0,833	0,671	0,657	4	6
agenda control_de	zum eigentlichen Thema	4	0,75	0,75	0	-0,143	3	0
agenda control_de	du wiederholst dich	9	0,889	0,889	0	-0,059	8	0
agenda control_de	bleib beim Thema	16	0,688	0,688	-0,148	-0,185	11	0
agenda control_en	come back to	10	0,9	0,9	0,808	0,798	5	4
agenda control_en	back to topic	13	0,846	0,846	0,653	0,639	8	3
agenda control_en	stick to topic	16	0,938	0,938	0,644	0,632	14	1
agenda control_en	focus on topic	4	0,75	0,75	0,533	0,467	1	2
agenda control_en	back to the topic	16	0,875	0,875	0,446	0,429	13	1
agenda control_en	stay on topic	41	0,732	0,732	0,375	0,367	7	23
agenda control_en	focus on issue	45	0,933	0,933	0,372	0,365	1	41
agenda control_en	focus on the topic	9	0,556	0,556	0,056	0	4	1
agenda control_en	whataboutism	62	0,774	0,774	0,003	-0,005	1	47
agenda control_en	arguments presented	19	0,947	0,947	0	-0,027	0	18
agenda control_en	get back to the main point	9	0,778	0,778	-0,062	-0,125	0	7
consensus seeking_de	guter Standpunkt	2	1	1	1	NA	0	2
consensus seeking_de	guter Punkt	23	0,957	0,957	0	-0,022	0	22
consensus seeking_de	darauf einigen	43	0,628	0,628	-0,214	-0,229	0	27
consensus seeking_en	common goal	10	1	1	1	NA	0	10

consensus seeking_en	can we agree that	22	0,955	0,955	0,834	0,83	3	18
consensus seeking_en	agree upon	6	0,833	0,833	0,593	0,556	1	4
consensus seeking_en	valid point	79	0,899	0,899	0,545	0,542	6	65
consensus seeking_en	agree with me	27	0,926	0,926	0,47	0,46	1	24
consensus seeking_en	can you agree with me	2	0,5	0,5	0	-0,333	0	1
consensus seeking_en	can you understand	23	0,826	0,826	-0,071	-0,095	0	19
consensus seeking_en	can we agree upon	3	0	0	-0,667	-1	NA	NA
engaging participation_de	näher erläutern	22	0,955	0,955	0,863	0,86	17	4
engaging participation_de	weiter erklären	18	0,944	0,944	0,646	0,636	1	16
engaging participation_de	Meinung hören	19	0,895	0,895	0,615	0,604	2	15
engaging participation_de	deiner Ansicht nach	38	0,816	0,816	0,129	0,118	1	30
engaging participation_de	neugierig sein	27	0,963	0,963	0	-0,019	0	26
engaging participation_en	what was your intention	4	1	1	1	NA	0	4
engaging participation_en	what do you intend	6	1	1	1	NA	0	6
engaging participation_en	open to feedback	12	1	1	1	1	1	11
engaging participation_en	open to criticism	5	1	1	1	NA	0	5
engaging participation_en	can you respond	2	1	1	1	NA	0	2
engaging participation_en	hear your view	36	0,972	0,972	0,945	0,944	16	19
engaging participation_en	explain to me	16	0,938	0,938	0,874	0,87	6	9
engaging participation_en	explain further	21	0,952	0,952	0,778	0,773	2	18
engaging participation_en	can you elaborate	21	0,952	0,952	0,778	0,773	18	2

engaging participation_en	curious to hear	10	0,9	0,9	0,747	0,733	2	7
engaging participation_en	constructive criticism	43	0,721	0,721	0,167	0,157	3	28
engaging participation_en	you mean	4	0,75	0,75	0	-0,143	0	3
engaging participation_en	what did you intend	7	0,857	0,857	0	-0,077	0	6
engaging participation_en	explain your argument	7	0,857	0,857	0	-0,077	0	6
engaging participation_en	could you respond	12	0,917	0,917	0	-0,043	0	11
engaging participation_en	would like to understand	19	0,789	0,789	-0,088	-0,118	0	15
engaging participation_en	interested in your	6	0,667	0,667	-0,1	-0,2	0	4
engaging participation_en	open for criticism	14	0,714	0,714	-0,125	-0,167	0	10
engaging participation_en	interested in your opinion	25	0,72	0,72	-0,14	-0,163	0	18
engaging participation_en	what is your intent	6	0,5	0,5	-0,222	-0,333	0	3
engaging participation_en	happy to discuss	26	0,577	0,577	-0,244	-0,268	0	15
engaging participation_en	interested in your view	22	0,455	0,455	-0,344	-0,375	0	10
engaging participation_en	interested to hear	49	0,408	0,408	-0,406	-0,42	0	20
engaging participation_en	what is your intention	7	0,286	0,286	-0,444	-0,556	0	2
engaging participation_en	can you explain further	19	0	0	-0,947	-1	NA	NA
showing openness_en	open to critique	22	1	1	1	NA	0	22
tone policing_de	freundlich bleiben	11	1	1	1	1	2	9
tone policing_de	kein Fass aufmachen	15	0,933	0,933	0	-0,034	0	14
tone policing_de	beruhig dich	20	0,6	0,6	-0,04	-0,067	1	11
tone policing_de	entspann dich	15	0,4	0,4	-0,249	-0,292	5	1

tone policing_en	watch your tone	3	1	1	1	NA	0	3
tone policing_en	stay friendly	28	1	1	1	NA	0	28
tone policing_en	don´t insult me	10	0,9	0,9	0,627	0,608	1	8
tone policing_en	calm down	24	0,667	0,667	0,13	0,111	2	14
tone policing_en	listen to each other	16	0,938	0,938	0	-0,032	0	15
tone policing_en	take it personal	75	0,947	0,947	-0,021	-0,027	0	71

Appendix 3: Discussion of Computational Sampling instead of Phrases

The research methodology employed to sample users moderating in social media deserves a short explanation: originally, the approach was to use computational methods to search social media for existing cases of user moderation and use these inductively to develop the categories for different moderation strategies. Using a combination of sentiment analysis, topic stringency, argument mining and author network analysis yielded a set of moderating statements from which only a small percentage (<.001 %) could be labeled as partially moderating with an intercoder reliability score of higher than .7.

Producing a sizable corpus this way would not have been feasible. The approach was abandoned, when it became clear that an even smaller percentage of these cases belonged to reply-chains on Twitter that fulfilled our requirement of five or more subsequent posts.

This research was conceived and carried out before ChatGPT came out. To avoid confirmation bias, the initial computational approach might be worth another attempt, having these capabilities at the disposal instead of the phrase-based sampling. Nevertheless, the decidedly longer time of hand-coding and qualitative analysis led to some important insights into the gray areas of the developed definition of moderation, namely the question of how real the users' intent to moderate is.

Another advantage of using inductive sampling would be to learn something about the real distribution of moderation strategies. An automated approach could make the case of analyzing enough conversations that they could be thought of as a sample for the entirety of Twitter. Based on these results, and given the arguments presented earlier, a general trend towards a skewed distribution containing fewer high-level strategies can be observed.

Appendix 4: Labeling Issue-Focused Reactions to User 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 emerg**e** 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. (I)*
    - Differentiation: Highlighting the differences between two or more concepts, ideas, or perspectives. (I)
    - Validation: Providing support or evidence to confirm the validity or correctness of a statement or argument. (I)
    - Ratification: Officially approving or confirming a decision, action, or agreement. (I)
2. Negative Counter-horizons:
    - Antithesis: Presenting a contrasting or opposing idea or concept. (I)
    - Opposition: Expressing disagreement or resistance to a particular viewpoint or proposition. (NI)
    - Divergence: Discussing or exploring alternative paths or possibilities that deviate from the current direction or consensus. (NI)
3. Closing Thematic Conclusion:
    - Positive Evaluation: Offering a positive assessment or judgment of the overall theme, topic, or argument. (I)
    - Opinion Synthesis: Bringing together different opinions, perspectives, or arguments to form a cohesive and balanced conclusion. (I)
4. Ritual Conclusion:
    - Change of topic: Shifting the discussion to a different subject or area of focus. (NI)
    - Formal synthesis: Summarizing the main points or ideas discussed in a structured and organized manner. (NI)
    - Rejecting the topic: Explicitly dismissing or refusing to engage with the current topic or proposition. (NI)

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. (NI)

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. (I)

* For the summary in the paper, these categories were grouped under issue-focussed/consensus seeking (I) or not (N). A more differentiated analysis including the role of emotions [TODO list references] and deliberation is planned. Whether positive emotions are a sign of deliberative quality is unclear. We posit that they don't hurt the process at least.

Appendix 5: Graphical Overview over different Phases of the Research