Guidelines for User Moderation Annotation

User Moderation Annotation: Goal

Moderation in most platforms relies generally on *expert moderators*, who are trained specifically for the role and whose contribution in the platform most often is specifically that of moderation. However, the role of moderation in deliberation and argumentation platforms can often be seen in general *user comments*; and the impact or contribution of that comment to the discussion is in line with that of a moderator. The goal of this study is to annotate *user comments* that align with the characteristics of [expert] moderation within a discussion or argument.

What is Moderation?[1]

The goal of *moderation* in deliberation and argumentation platforms is to create an environment of informed and thoughtful participation, as well as mentor effective commenting behavior.

A moderator moves participants past "voting and venting" behaviors to effectively contributing the information they possess. They also make participants feel that their voices have been heard and that they are part of a forum for [civil] engagement.

Moderators have the role of advocating for the commenting process; as they encourage a "knowledge building community" that supports commenters' access to, participation in, and learning about the process and topic under discussion. Whether the goal of the process is policymaking, converging perspectives, or arguing one's view, moderation helps commenters to contribute as individuals as well as collaborate with each other.

Expectation of Moderators

1. Neutrality: Expert moderators are strongly encouraged to remain *neutral*, avoiding taking a position on the substance of the discussion, or forming biases or making assumptions about participants' comments. However, users are not restricted to this requirement and comments that do indeed have the role of moderation from a user may (e.g. in the case of clarification comments) or may not (e.g. signaling erred information to another user) have this characteristic.
2. Maintaining the norms: Expert moderators are responsible for maintaining the norms of the platform community and its regulations. Users might mirror this role in subtle ways, such as reminding others of the goal of the discussion or pointing out inappropriate contributions.
3. Choice of wording: Expert moderators are asked to use plain language, calm tones, avoid condescending responses, and limit the number of questions. For example:
    a. *That clarification is available in several forms on the website http:[...]*
    b. *DOT has estimated that the benefits of this discussion will outweigh the costs.*
    c. *This is an interesting suggestion, thanks. Could you provide a little more information on this, and perhaps a link.*

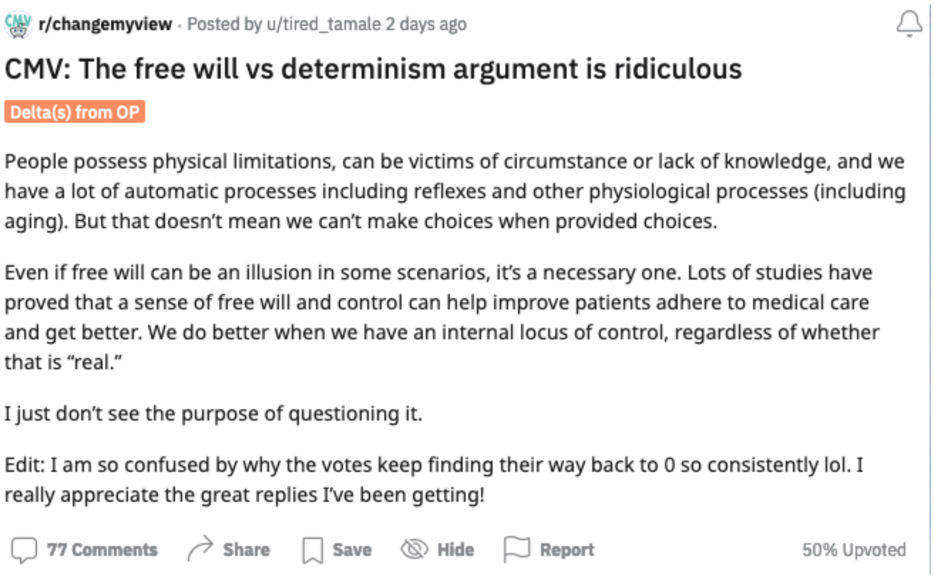
    Again, users are not expected to uphold these standards in their comments, however they may still perform similar contributions to the discussion, with or without a careful choice of wording.

[1] Moderation overview is adapted from the Moderator Protocol of RegulationRoom.com

The Data: Change My View

The data you will be annotating is extracted from the online subreddit entitled Change My View.[2] The platform is dedicated to civil discourse, aimed at promoting productive conversation to resolve differences by understanding others' perspectives.

The format of CMV is as follows. First, a user (original poster, or OP) posts a *view*, defined as a particular way of considering or regarding something, an attitude or opinion, on a specified *topic* issue, and asks the community to "change my view". For example:



r/changemyview · Posted by u/tired_tamale 2 days ago

CMV: The free will vs determinism argument is ridiculous

Delta(s) from OP

People possess physical limitations, can be victims of circumstance or lack of knowledge, and we have a lot of automatic processes including reflexes and other physiological processes (including aging). But that doesn't mean we can't make choices when provided choices.

Even if free will can be an illusion in some scenarios, it's a necessary one. Lots of studies have proved that a sense of free will and control can help improve patients adhere to medical care and get better. We do better when we have an internal locus of control, regardless of whether that is "real."

I just don't see the purpose of questioning it.

Edit: I am so confused by why the votes keep finding their way back to 0 so consistently lol. I really appreciate the great replies I've been getting!

Users are then able to interact with the OP as comments to argue their perspective in order to change the OP author's view. The interaction between users and OP author may be a simple back-and-forth comment, or may be an extended discussion. At the end of the interaction, if the user's argument has successfully changed the OP's view, the user is awarded a *Delta* (Δ) by the OP author.

Annotation Task

The annotator will be shown two texts: the ***preceding comment*** (for example, the OP or a post in the comment thread) and the ***reply comment***. The preceding comment as well as the topic of the OP are provided to the annotator to offer context. The *reply comment* is the comment to which the annotation questions refer. For each reply comment, the annotators are asked a set of questions, described in detail below:

1. **User Moderation** [y/n]

   Do you consider this user comment to behave as a form of moderation in the discussion?

[2] https://www.reddit.com/r/changemyview/

2. **Moderation Function**[3]

   In the case that the user comment behaves as a form of moderation, please provide information on the type of *moderation function* the comment performs. Please select the most appropriate function(s), understanding that the language use of users may lead to more flexibility and interpretation of the definitions of these moderation functions. After selecting the relevant functions, the annotators may provide additional comments or justification for their selection as a short answer.

   a. **Broadening Discussion**. [y/n] The comment encourages users to consider and engage comment of other users; or it promotes a more expansive or broader discussion on the topic by the author of the preceding comment or the community.

   b. **Improving Comment Quality**. [y/n] The comment asks for more information, factual details, or data to be provided to support the statements made; or asks the author of the preceding comment to make or consider possible solutions or alternative approaches.

   c. **Content Correction**. [y/n] The user comment provides substantive information about the preceding comment; corrects misstatements or clarifies details about the preceding comment; or points to relevant information such as websites or specific documents with the goal of correcting the content of the preceding comment.

   d. **Keeping Discussion on Topic**. [y/n] The user comment explains why the preceding comment is beyond the authority or competence of the platform, or outside the scope of the discussion; or it indicates irrelevant, off-point statements.

   e. **Organizing Discussion**. [y/n] The comment directs the author of the preceding comment to another post or comment that is more relevant to their expressed interest.

   f. **Policing**. [y/n] The comment aims to maintain/encourage civil deliberative discourse; or it points to inappropriate language or content in the preceding comment.

   g. **Resolving Site Use Issues**. [y/n] The comment is to resolve technical difficulties; or it provides information about the goals/rules of the platform.

   h. **Social Functions.** [y/n] The user comment takes on the function of welcoming/greeting, encouragement or appreciation of the preceding comment, or thanking for participation.

3. **Justification** (Optional)

   You can provide a short justification or any details you would like to offer for your answers to questions (1) and (2). Please note, there is a limit of 225 characters for this answer.

4. **Constructiveness** [1-5 scale]

   Considering the user comment in general – whether or not it behaves as a form of moderation – do you consider this comment to be constructive to the discussion?

   Constructive comments can be defined as *high-quality comments that make a contribution to the conversation*. Such comments are considered to offer an opinion or perspective, and provide support, reasoning, or background for that view. They are characterized as comments that intend to create a civil

[3] Taken from Moderator Roles and Interventions (Park et al., 2012)

dialogue through remarks that are relevant to the discussion/topic and not intended to merely provoke an emotional response.

5. **Sentiment / Tone** [positive | neutral | negative]
   How would you evaluate the overall tone of the user comment? Would you consider the underlying feeling, attitude, evaluation, or emotion associated to the comment as positive, negative, or neutral?

6. **Subjectivity** [1-5 scale]
   Does the user comment refer to the user's personal opinions or feelings regarding a particular subject matter, based on their unique interpretation of an idea or their own thoughts, feelings, and background; or is the comment rather neutral in this respect?

7. **Aggressivity** [1-5 scale]
   Do you consider the user comment to be aggressive, actively or passively? Examples could include (but are not limited to) sarcasm, blaming, intimidation, threats, or attacks.

8. **Agreement with comment opinion** [yes | no | opinion not clear]
   Do you agree with the opinion expressed in the reply comment?

Trigger Warning!

As mentioned in the consent form you agreed to, the texts included in this study are produced in an online debate forum and some topics that are discussed, how they are discussed, and user perspectives may be uncomfortable or sensitive. First, all texts included do not represent the views of the researchers conducting the study. Secondly, we provide the option to avoid having to annotate any instance that is problematic or uncomfortable for the annotator without penalty of compensation.

To do so, please answer the annotation questions as outlined below. Note, although you will have provided answers, if you include the following text in the Justification, your answers to this instance will be automatically discarded and not considered in the study.

1. User Moderation: *No*
2. Moderation Function: *None of these*
3. Justification: (please copy and paste) *I am uncomfortable annotating this text and voluntarily skip this instance.*
4. Constructiveness: *No*
5. Sentiment: *Neutral*
6. Subjectivity: *Neutral*
7. Aggressiveness: *Neutral*
8. Do you agree with opinion: *Opinion not clear*