Supplementary materials to *How to co-create content moderation policies: The case of the AutSPACEs project*

S1: Details on the co-creative processes outlined in the results paper

Due to length constraints in the main body of the manuscript, we could only give a high-level summary of the co-creation process and its impact. Here, we give a selection of more detailed examples of how the iterative co-creation steps improved the different content moderation aspects. In particular, we document the individual data collected to demonstrate these different aspects, to make the process as a whole more transparent.

Process

1. **Scoping sessions**: during the scoping sessions it emerged that there was a desire for an online space to share experiences, and that sensory processing was a highly important topic which had impacted people greatly. A diversity of views were expressed about the benefits and drawbacks of sharing their sensory experiences with other autistic people compared to researchers.
2. **Focus Groups:** During the focus groups, the topic of moderation drew much discussion. There was passionate disagreement about what content would be appropriate to share, and what effect reading the comments of others could have on platform users. Researchers also shared their own views as part of the discussion. These focus groups were transcribed and then the transcription was turned into an anonymised version with comments coded according to who was speaking and their relationship to the experience or observation they

were sharing and thematically grouped. The resulting document is published on GitHub (see links above)

3. **Moderation workshop:** This workshop centred around the open questions of how content should be moderated. Similarly to the focus groups, the workshop was transcribed and an anonymized version of annotated quotes uploaded publicly on GitHub (see link above).
4. Tasks/Priorities around moderation were identified and turned into public issues on GitHub. These were labelled with a specific "moderation" tag. (c.f. https://github.com/alan-turing-institute/AutisticaCitizenScience/issues?q=is%3Aissue+is%3Aclosed+milestone%3AModeration for the corresponding issues).
5. In parallel, a risk register was created by one of the autistic co-leads to track possible risks and begin to strategise options to mitigate them: https://github.com/alan-turing-institute/AutisticaCitizenScience/blob/master/Moderation/Moderation%20Strategic%20thinking.docx
6. The core moderation group met to co-work on these issues, identifying dilemmas and potential solutions
7. Potential solutions are presented to the larger community at the regular online meetups for feedback, improvements etc.

Based on this overall workflow, we present some detailed examples of how the different dilemmas and potential solutions were identified and refined by the community.

Manuscript Section 3.2.1. No commenting on other users posts

1. Based on the initial focus groups and workshop data, the initial design decision is made to not allow for comments (see main text of the manuscript).
2. This decision established early in the project was maintained as evidence from focus groups suggested that the risks out-weighed the benefits of allowing comments, and the duty to mitigate potential harm to participants was considered paramount by researchers.
3. The decision was finalised and set into the moderation process design, with the option to iterate following the platform launch.

Manuscript Section 3.2.2. Content submission and moderation guidelines with rules for content warnings

1. During the focus group sessions, the possibility of self-moderating what comments could be seen was proposed by an autistic participant, while another autistic participant suggested colour-coding for different kinds of experience, see quotes below:

   "(A d): include option to moderate out distressing or negative experiences: "if I'm feeling particularly stressed and I'm going to go on a platform, I don't want to be reading…a lot of what went wrong on your day, I'd rather click on a button that filters to the – what happened, and what went well, and who accommodated you well"

    - (A d): different types of experience could be colour-coded" (https://github.com/alan-turing-institute/AutSPACEs/blob/main/00-project-documentation/community/focus-groups/18-september-2019.md#self-moderation)
2. The core team on moderation jointly developed an initial "traffic lights" system
3. The system was presented to the community during meet-up sessions on the 24th March, the 22nd April 2021, and the 27th May 2021, and changed on the basis of their feedback. In particular, the initial idea of having a penalty of fully "blocking" contributors if they break the guidelines was removed as community members pointed out that it might not be a case of malicious posting so much as misunderstanding, and that the approach was therefore overly punitive.

Manuscript Section 3.2.3. Guidelines for sharing stories about others

1. Based on the importance of the topic during the focus groups & scoping sessions, the question of whether parents and carers of autistic people should be allowed to use the platform and if so on what basis was a large part of the moderation co-creation workshop. Direct quotations from the focus groups, illustrating the range of views and reasons for them, were presented to the attendees of the workshop. A mix of autistic people, non-autistic parents of autistic people, and those who were both autistic themselves and parents of autistic people were included in the workshop in order to garner views from across these different groups.
2. After the worksop, the option of having guidelines for parents was suggested by an autistic moderation co-lead and Initial guidelines were written up collaboratively by a researcher and the autistic co-leads
3. These guidelines were then discussed with the larger community during online meetups (on 2nd February 2023), and following their input adjustments were made to the guidelines, including providing clear examples added to illustrate the difference between acceptable and unacceptable versions of an experience shared by a parent of an autistic person to improve clarity. Implementation within the platform of ways of distinguishing between direct and indirect reports of experiences were also discussed with the community and tested on 7th September 2023.

Manuscript Section 3.2.4. When to moderate: pre-publication review

1. During the focus groups & moderation workshop, many autistic participants shared their experiences of being the target of abuse or witnessing abuse online. During the workshop, an initial consensus was reached for pre-publication moderation despite recognizing the possible risk of causing frustration or putting people off using the platform.
2. A time scale for moderation was proposed which was discussed with autistic people. During this process, it was found that it was clear to users to expect a delay following the submission of an experience for publication due to the need for moderation. Such a delay was deemed acceptable by community members and worthwhile if it allowed for a more thorough vetting of posts and reduced the risk of damaging or anti-autistic messages being posted.

3. The “moderation flow” for how users and moderators would interact with the platform was developed by the core moderation team, and then the design in the form of a flow chart was presented to the larger community in a number of meet-up sessions. This flow was refined with the input of autistic collaborators in both online co-working sessions and meet-up sessions.
4. In particular, as a result of this larger community conversation, it was decided to not allow moderators to write up open responses on why a submission was rejected, but rather provide a closed list that matches the sections of the content moderation guidelines from which moderators can choose (see community meetup notes from 2nd February 2023). This means submitters have information on why their post was denied for publication, but it is not overly taxing on the time resources of moderators and it remains as objective as possible.

S2: Links to data

The data from the focus groups, workshops and regularly occurring community meetings have been publicly archived in the GitHub repository of the project alongside documentation of how the data was collected, annotated etc.

Focus groups

The data and supporting/explanatory materials can be found at https://github.com/alan-turing-institute/AutSPACEs/tree/main/00-project-documentation/community/focus-groups

The main README file explains the data itself, as well as the metadata keys. Furthermore, the “creation of summaries- file explains in detail how each of those datasets was created from how the focus groups were run, how the data was transcribed and annotated to the final data set, this file can be found at https://github.com/alan-turing-institute/AutSPACEs/blob/main/00-project-documentation/community/focus-groups/creation-of-summaries.md

Moderation design workshop

Analogous to the focus group data, we provide the annotated data set for the moderation workshop alongside the documentation of how the workshop was run on GitHub at https://github.com/alan-turing-institute/AutSPACEs/tree/main/00-project-documentation/community/moderation-workshop

The data in this report are the outcome of the group discussion, which was first transcribed by a researcher, while excluding personal identifiable information, and then collecting the anonymised quotes, before being reviewed by the participants following the process outlined in this data review document:

https://github.com/alan-turing-institute/AutSPACEs/blob/main/00-project-documentation/moderation/moderation-review-guidelines.md (This explanatory text for how the data was processed was also co-produced with the inclusion of an autistic community member)

Content moderation guidelines

The final content moderation guidelines, including the documentation for moderators can be found on GitHub at https://github.com/alan-turing-institute/AutSPACEs/blob/main/00-project-documentation/moderation/moderator-guidelines.md

Meet-up sessions

For the open community meet-up sessions notes were taken collectively on Google Docs. The notes are available on the projects Wiki page: https://github.com/alan-turing-institute/AutSPACEs/wiki/Meetups