Data & Policy (2024), 6: e28
doi:10.1017/dap.2024.21

TRANSLATIONAL ARTICLE

How to co-create content moderation policies: the case of the AutSPACEs project

Georgia Aitkenhead[1], Susanna Fantoni[2], James Scott[2], Sophia Batchelor[1], Helen Duncan[1], David Llewellyn-Jones[1], Callum Mole[1], Otis Smith[1], Martin Stoffel[1], Robin Taylor[2], Kirstie Whitaker[1] and Bastian Greshake Tzovaras[1]

[1]The Alan Turing Institute, London, UK
[2]Citizen Science Contributors
Corresponding author: Bastian Greshake Tzovaras; Email: bgreshaketzovaras@turing.ac.uk

G.A., S.F., and J.S. have contributed equally to this work.

Received: 11 July 2023; **Revised:** 16 February 2024; **Accepted:** 16 April 2024

Keywords: autism; citizen science; co-design; content moderation; social networks

Abstract

The moderation of user-generated content on online platforms remains a key solution to protecting people online, but also remains a perpetual challenge as the appropriateness of content moderation guidelines depends on the online community that they aim to govern. This challenge affects marginalized groups in particular, as they more frequently experience online abuse but also end up falsely being the target of content-moderation guidelines. While there have been calls for democratic, community-moderation, there has so far been little research into how to implement such approaches. Here, we present the co-creation of content moderation strategies with the users of an online platform to address some of these challenges. Within the context of AutSPACEs—an online citizen science platform that aims to allow autistic people to share their own sensory processing experiences publicly—we used a community-based and participatory approach to co-design a content moderation solution that would fit the preferences, priorities, and needs of its autistic user community. We outline how this approach helped us discover context-specific moderation dilemmas around participant safety and well-being and how we addressed those. These trade-offs have resulted in a moderation design that differs from more general social networks in aspects such as how to contribute, when to moderate, and what to moderate. While these dilemmas, processes, and solutions are specific to the context of AutSPACEs, we highlight how the co-design approach itself could be applied and useful for other communities to uncover challenges and help other online spaces to embed safety and empowerment.

Policy Significance Statement

Content moderation in online communities is a hard problem that has no simple solutions. Given the different priorities and trade-offs that participants in online communities would prefer around the moderation of content, all content moderation guidelines have to end up being uneasy compromises around weighing different dilemmas. This work outlines how engaging in the co-creation of content moderation guidelines—with those who will be ultimately governed by them—can provide a pathway to uncover community-specific dilemmas and priorities. Additionally, it provides a way to make informed decisions when weighing potential solutions against each other. As such, we expect this work to be informative for policymakers working on implementing moderation policies themselves.

https://doi.org/10.1017/dap.2024.21 Published online by Cambridge University Press

1. Introduction

How to moderate content on online platforms remains a key question when it comes to protecting people online (Gillespie, 2017). In particular, these questions can affect marginalized groups that might be at a particularly high risk of both online abuse or be the victims of "false positives" where they are the target of moderation policies (Feuston et al., 2020; Haimson et al., 2021; Salty, 2021). Whilst there are a variety of approaches to content moderation—such as pre- or post-publication moderation, algorithmic moderation, and community moderation (Veglis, 2014)—there is limited research into which approaches work best under which circumstances, though in recent years there have been increasing calls for democratizing moderation (De Gregorio, 2020) and making use of community self-moderation (Seering, 2020).

The question of how content is moderated online is of particular relevance to people who are not falling within the neurotypical majority, including autistic people. It is generally known, that social spaces that are built by and for neurotypical individuals are often excluding to autistic people—in large part due to a lack of understanding and misperception by the neurotypical majority (Sasson et al., 2017; Morrison et al., 2019; Mitchell et al., 2021). The particular challenges that autistic people can encounter in online spaces as a result of such exclusionary design are an abuse of trust and being the target of cyberbullying. Studies show some evidence that autistic people's moral judgments may be more heavily influenced by the moral evaluations of the situation rather than the intentions behind the actions of others (Zalla et al., 2011), which in turn may hinder the ability to predict individual intentions (Chambon et al., 2017). Considering that online spaces can lower the barrier to acting deceptively, this may create a large concern and there is preliminary evidence showing that autistic people are at higher risk of sexual exploitation (Landon, 2016) as well as cyberbullying (Trundle et al., 2022). As it can be hard for autistic people to determine the intentions of their peers, this can create a power imbalance and raise the likelihood of bullying both amongst autistic children (Campbell et al., 2017) and autistic adults, who are at risk of being ignored or left out of online interactions, leading to an increased feeling of worthlessness and negativity (Triantafyllopoulou et al., 2022).

Challenges relating to content moderation can have substantial impact as online social spaces such as social media can be a source of support for autistic people, including better friendship quality in adolescents (van Schalkwyk et al., 2017), and closer relationships in adults (Ward et al., 2018). Furthermore, the structured format of social media sites might give autistic users the chance to contribute without having to assess how it is being perceived (van Schalkwyk et al., 2017). More recent work focusing on teenagers has emphasized these points (Gillespie-Smith et al., 2021): Participants outlined how online interactions let them avoid the stress of face-to-face discussions with the screen acting as a shield and protection that affords a sense of security that enables the discussion of topics that would be hard otherwise. Given that it has been suggested autistic people might be spending more time online than their non-autistic peers (Wang et al., 2020), both the negative and positive impacts of social media use are likely to be amplified amongst this population.

While governance questions around content moderation are most commonly asked in the context of social media platforms, they also apply to other online systems, including the ones used for citizen science. Citizen science describes a broad set of methods and practices that all center around involving non-researchers in the act of doing research (Haklay et al., 2021), including in data collection, data processing but also in the co-design of research studies (Vohland et al., 2021). While historically often used in fields such as ecology or astronomy, citizen science is increasingly used in health-related research too (Remmers et al., 2023). While content moderation questions in citizen science are mostly framed around quality and safety issues of citizen-collected data (Kapenekakis and Chorianopoulos, 2017; Schacher et al., 2023), how these online platforms are designed and governed are also actively studied (Kloppenborg et al., 2021; Morell et al., 2021).

Here, we present how we co-designed and co-created a content moderation approach for an online citizen science research project named AutSPACEs—short for *Autism Research into Sensory Processing for Accessible Community Environments*—that investigates sensory processing differences experienced by autistic people, and how this affects them in their daily life. Overall, this raised the question of how our online platform design and implementation can encourage the positive outcomes for autistic people that we outlined above, without exposing users to the negatives?

https://doi.org/10.1017/dap.2024.21 Published online by Cambridge University Press

Using AutSPACEs as a case study, we highlight how we used participatory design and research methods (Spinuzzi, 2005; Senabre Hidalgo et al., 2018) to collaboratively collect and analyze data to uncover particular content moderation requirements and then co-design moderation approaches that address these. Our goal is to use this case study to demonstrate how co-design can be a powerful tool to create content moderation policies that are rooted in evidence-based user requirements to provide other online platform creators—in citizen science and beyond—with an approach that they adapt to co-develop their own content moderation approaches. To that end, we first briefly introduce the background of the AutSPACEs project as well as the co-design methods we used. Then we show the particular content moderation dilemmas our collaborative analysis uncovered for AutSPACEs, followed by the co-designed solutions to try to address them. Lastly, we discuss the implications of both the particular content moderation approach as well as the co-design strategy.

1.1. Our case study: the AutSPACEs project

AutSPACEs is a citizen science project that was started in 2019 with the aim of gathering qualitative data at scale to investigate how sensory processing sensitivities affect autistic people's navigation of the world for a neurodiverse community of citizen scientists through a welcoming and inclusive online platform. Research on sensory processing and autism has shown that sensory processing sensitivities are extremely prevalent (Ben-Sasson et al., 2009; Crane et al., 2009) and can have significant effects on the lives of autistic people and their families (Schaaf et al., 2011; Fletcher-Watson and Happé, 2019). So far, sensory processing and its impact on people's daily lives (e.g., in schools, workplaces, or hospitals) are not well understood, which is why the interrelated questions of "How can sensory processing in autism be better understood?" and "What environments/supports are most appropriate in terms of achieving the best education/life/social skills outcomes in autistic people?" emerged as key research priorities for autistic people in the James Lind Alliance priority-setting partnership (Cusack and Sterry, 2016).

AutSPACEs aims to help address this gap by gathering data from these lived experiences for research but also to help educate non-autistic people about these challenges and improve the design of spaces. These experiences are contributed as personal stories in text form, responding to two free-text prompts: "What was your sensory processing difference?" and "What could have made your experience better?" While these experiences can be privately entered and be flagged for internal research use, users of AutSPACEs are also given the option to publicly share their experiences through the platform, thus giving others the chance to read and learn from them but also requiring the creation of content moderation guidelines.

AutSPACEs can be understood as a case of "extreme citizen science," a practice in which participants not only contribute data but also control the design and implementation of the research (Skarlatidou et al., 2022). As such, AutSPACEs is facilitated with the help of a wide range of institutional partners (The Alan Turing Institute; Autistica; Open Humans) as well as a growing community of autistic contributors, open source developers, and organizations—such as Fujitsu and the UK Civil Service—that have made contributions. This collaborative working across a wide set of multiple stakeholders is core to AutSPACEs, as is its deliberative research process that is co-led by autistic citizen scientists. This approach is also reflected amongst the authors of this contribution, as the team that developed the moderation approach consists of a mix of viewpoints, including autistic/non-autistic authors as well as academic and citizen scientists. In recognition of historic and systemic power imbalances, we have developed this participatory, inclusive, and community-led stance to center autistic people and their lived experiences in our research, providing a platform for autistic people to speak for themselves.

2. Methods

The co-design of the content moderation for AutSPACEs was done using a number of different participatory strategies that involved the larger community in different ways, as shown in Figure 1 In the initial stages, scoping sessions (a) as well as focus groups and workshops (b) were used to collect transcribed data (c). Following this, a more iterative co-design stage took place: A moderation core team consisting of researchers and autistic contributors analyzed the data for moderation requirements (d). Based on these analyses, the team developed moderation strategies (e) that were presented for the larger community (f) to then further refine the moderation approaches. Ultimately, this culminated in the final content moderation guidelines for AutSPACEs (g). This approach has now been implemented into the technical design of the AutSPACEs platform. We are currently finalizing the platform development, with the launch of it expected to happen in 2024.

Below, we give a brief summary and overview of our co-creation approach. Supplementary Material gives a more detailed description of the co-creation approach (S1), as well as direct links to the full documentation (S2).

2.1. Scoping, focus groups, and workshops

2.1.1. Scoping

In the early stages of the project, two scoping sessions were held at The Alan Turing Institute, comprising a total of 26 participants. The first consisted of 13 people in a single group discussion, while the second session was broken into two groups. Both sessions were made up of autistic people and parents and carers of autistic people (some intersected both categories) who were based in the London area. These participants were deliberately chosen to be as wide a demographic as possible, with both meetings having delegates from a wide range of age brackets (ranging from delegates aged 18–25 to including delegates aged 65+), as well as a mix of women and men. The goal of the scoping sessions was to derive principles for the research process (such as data management and study protocols) and to identify priorities and challenges for the project with the input of autistic people, their families, and carers. At this early stage, we kept the permissions we sought from participants for sharing data to a minimum, instead emphasizing privacy, until having come to a participatory consensus on the appropriate handling of data with the community. Data from the scoping phase thus remains private and is held by Autistica and only high-level summaries are presented in this article.

2.1.2. Focus groups

Based on these scoping sessions, we developed a more participation-focused process for the co-design phase, as well as a more precise idea of the project concept and formulated these aspects for institutional

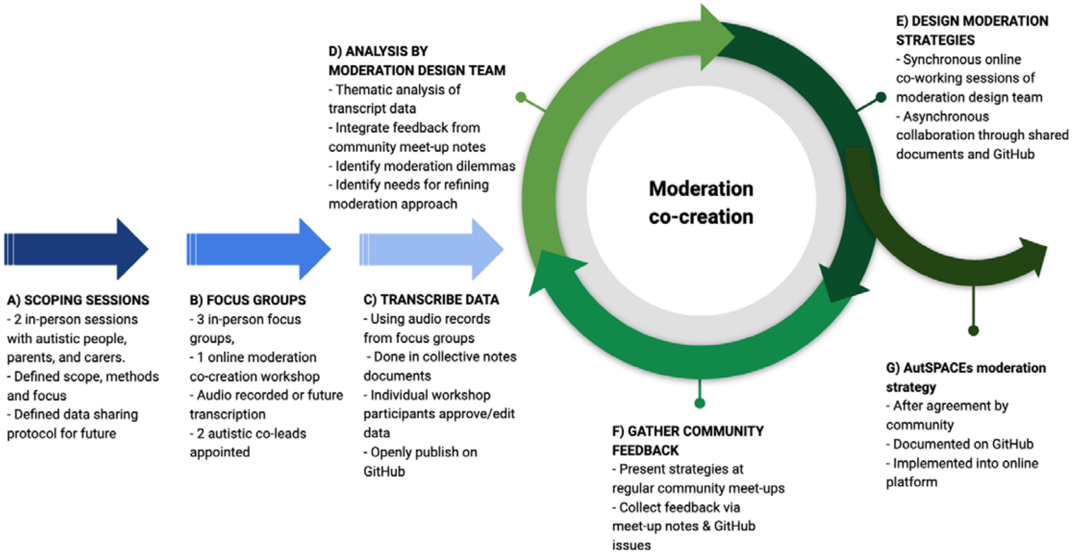
Moderation co-creation
A) SCOPING SESSIONS
- 2 in-person sessions with autistic people, parents, and carers.
- Defined scope, methods and focus
- Defined data sharing protocol for future
B) FOCUS GROUPS
- 3 in-person focus groups,
- 1 online moderation co-creation workshop
- Audio recorded or future transcription
- 2 autistic co-leads appointed
C) TRANSCRIBE DATA
- Using audio records from focus groups
- Done in collective notes documents
- Individual workshop participants approve/edit data
- Openly publish on GitHub
D) ANALYSIS BY MODERATION DESIGN TEAM
- Thematic analysis of transcript data
- Integrate feedback from community meet-up notes
- Identify moderation dilemmas
- Identify needs for refining moderation approach
E) DESIGN MODERATION STRATEGIES
- Synchronous online co-working sessions of moderation design team
- Asynchronous collaboration through shared documents and GitHub
F) GATHER COMMUNITY FEEDBACK
- Present strategies at regular community meet-ups
- Collect feedback via meet-up notes & GitHub issues
G) AutSPACEs moderation strategy
- After agreement by community
- Documented on GitHub
- Implemented into online platform

Figure 1. *The different stages of the moderation co-creation for AutSPACEs.*

https://doi.org/10.1017/dap.2024.21 Published online by Cambridge University Press

ethics approval. As a result, three in-person focus groups were then held with autistic community members, family members, carers and supporters of autistic people, developers, and researchers. In these sessions, moderation emerged as a topic which was viewed as a priority issue and which was furthermore especially contentious and so required original solutions and further deliberation.

2.1.3. *Moderation workshop*

As a result, we held an online workshop (held via Zoom), to target the questions of moderation more specifically and began the process of co-creating a moderation strategy for AutSPACEs. Participants in that session were presented with existing codes of conduct—from The Carpentries (2019), The Turing Way Community (2022), and AutAngel (2019)—as well as a range of comments about moderation from previous sessions to support the discussion of different moderation strategies. For all four of these sessions, the audio was transcribed and subsequently summarized to provide anonymity to participants. All participants consented to those anonymized data being published under an open license alongside all resulting outputs.

All data are available on GitHub at https://github.com/alan-turing-institute/AutSPACEs/. Direct "deep" links to each data set can be found in Supplementary Material S2.

2.2. *Moderation co-design*

Following the different focus groups and the moderation workshop, two neuroatypical volunteers—co-authors S.F. and J.S.—were appointed to co-lead on the moderation strategy by the research team. This led to an ongoing collaboration on moderation which took place online through regular virtual meetings, and asynchronous work in shared online documents and on the project's GitHub repository.

Our moderation core team started its design process through an (informal) thematic analysis, using the anonymized and annotated transcripts of the workshop session and focus groups as well as their own lived experiences and participant observations, as all members of the moderation team were part of the focus groups and workshops. This process led to the identification of a number of dilemmas in which opposing community preferences were identified that needed to be resolved or which at least informed decisions would need to be made.

Jointly, the moderation team consisting of the neurodiverse co-leads and the research team went through these conflicting wishes and tried to weigh them and make suggestions for resolving them. In an iterative process, these suggestions were then broad back to the larger AutSPACEs community for discussion in regularly occurring online community meetings that started in 2021 and are ongoing as of January 2024. Between 2021 and 2022 these meetings happened roughly every 2 weeks, since January 2023 they have switched to a monthly frequency. Based on the input of the larger community, the suggested approaches were then refined and shared again in later meetings. The meeting notes of these larger community discussions, as well as the final online content moderation guidelines are also available online in the project documentation folder on GitHub at https://github.com/alan-turing-institute/AutSPACEs/.

3. Results

The results of this work are twofold, and the results section is split into two corresponding sections as a case study of how we co-developed our content moderation strategy: In the first part, we examine a few of the main moderation-related dilemmas that our moderation team discovered through the thematic analysis of the transcripts of the scoping sessions and focus groups. In the second part, we highlight the decisions made around those dilemmas as well as how they were made.

3.1. *Moderation-related dilemmas*

Throughout our initial scoping sessions and the two focus groups, questions around who should be able to use the platform, who should moderate it, what content should be permitted to be published, and how we should manage that content and its moderation, emerged as critical points for the AutSPACEs community. Collectively, this made this a test case for approaching community collaboration for areas of high importance and low initial agreement. The core moderation team, which was formed as a response, did then analyze the collected records of the workshops and focus groups to identify shared key themes using this data in addition to their own participant observations. Through this process, multiple dilemmas related to ensuring the safety of AutSPACEs participants and content moderation emerged. We expand on some key dilemmas below, including giving exemplary statements related to them as voiced by participants in our workshops and focus groups.

3.1.1. Giving feedback on other's experiences

One contentious issue that our analysis uncovered was around whether participants should be able to reply to or comment on each other's experiences: Some participants in the focus groups thought that this would be a positive, as voiced by one autistic focus group participant: *I think that will encourage good will, because it will encourage people to [give] feedback more.* In contrast, two other autistic participants of the focus groups felt strongly about not wanting others to comment on their experiences, as their reports would be about *"my experience, there's nothing to debate about it,"* also worrying that *"there could be arguments"* as a result of allowing commenting. This led to consideration about whether a comment section should be included in the site or not and if so how that would be moderated in conjunction with the main content.

3.1.2. Potentially triggering content

We also found that a number of autistic participants shared the concern that reading about negative experiences could trigger distress. These concerns ranged from stressful experiences inducing or exacerbating stress in the reader (*"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"*), to even more severe consequences for those with mental health difficulties, such as being triggered by descriptions of abuse or suicidal behaviors (*"…there may be people who are writing about abuse, or are writing about suicide attempts. Those are then triggers for others"*). The community discussions also identified content that expresses stigma, prejudice, and discrimination, such as anti-autistic attitudes and behaviors as issues that present a trigger risk in the AutSPACEs system (*"in some of the dark…nastier corners of the internet they have a real thing about autism…using it as a slur against people, and taking the piss basically"*). In addition to causing distress, it was observed that the uncensored publication of negative stories about autistic people could perpetuate existing prejudices, undermining the purpose and values of the platform as those comments *"…can fuel the negative narrative, the deficit, the medical model… and that's the one thing that we can't do."*

As a general principle, the need to protect the mental well-being of the readers of public experiences from harmful content emerged (*"…we have got a duty of care"*). The necessity of moderation in online spaces was summarized by one autistic participant in particularly strong terms:

> I think you have to work on the basis, unfortunately, that every single corner of the internet which doesn't have moderation just seems to fill up with Nazis, they're everywhere, and it happens in the most unlikely places, so I would say at least for the first couple of times you need to moderate the users.

On the other hand, participants expressed the importance of autistic users being able to share negative or traumatic experiences: "*If those things do come up [around suicide or abuse], then there ought to be an opportunity to have people express what they're trying to say, but in a position where they're not… right on the edge of something dangerous.*" The benefit of both writing about and sharing difficult experiences was raised: "[*When] I write it helps me process, and every time I write something difficult, I write it with the mindset that if one person reads it and thinks they're not alone, or that they can cope better with their issues, then I achieved something,*" with another participant highlighting that they *"…think just people sharing their own story is very powerful."*

https://doi.org/10.1017/dap.2024.21 Published online by Cambridge University Press

Additionally, showing the genuine reality of autistic people was considered a benefit of the platform (*"awareness is really important—understanding"*). The benefits of knowing that others dealt with similar situations and of accessing shared strategies and solutions were highlighted, as "*…by doing that [posting strategies] and sharing with other people it helps autistic people, but also others who have relationships with them: so families, communities, schools…"*

In summary, a dilemma emerged consisting of a need to protect readers from being exposed to triggering or upsetting content, whilst allowing users to share their experiences publicly and to have these experiences accessible to those who might benefit from reading them.

3.1.3. Reporting on behalf of others

Arguably the biggest dilemma we discovered in our analysis was whether to allow someone to report experiences on behalf of others; for example, a parent reporting on behalf of an adult autistic child who may not be able to use the platform. In our early focus groups, there was a large diversity of opinions both from parents of autistic people and autistic people themselves. Many parents argued that reporting on behalf of their children was necessary, as their children's *"experiences would never be heard"* otherwise. Often this perspective was paired with a feeling that they *"should be able to share a story, because*" they felt as parents they would understand their children *"pretty well, and […] would hate for people's voices not to be heard because they can't express it."* Other parents saw a potential for a collaborative approach, with the power and decision-making ultimately resting with the autistic individual whose experience was being recorded: *"…my son very much is his own person, and he would want to go over what you've done anyway…so you'd never be able to get away with a wrong submission."*

Amongst many autistic participants, there was a very concrete concern that it would be "*very difficult to make a place welcoming to autistic people when you also have a lot of neurotypical people explaining about autistic people."* Thus, allowing reporting on behalf of others could potentially cause the site to become hostile and/or overwhelmed with non-autistic opinions. Some parents did recognize this danger and agreed that if reports by non-autistic people were allowed, it may cause *the platform [to become] less welcoming for autistic people.*"

Another concern expressed by autistic participants was that if reporting on behalf of others were to be allowed, this could lead to non-autistic parents using the site to air their frustrations. This was backed up by a fear of bias due to the fact that those reporting on behalf of others would be doing so from their perspective and not that of an autistic person and may *"exaggerate or post stories about other people."* The distortion introduced by this was viewed as particularly problematic, as autistic participants wanted *"as unbiased, authentic information as possible."* As one autistic participant put it: *"…we don't actually know the experience of this person because we don't have their testimony, we have the testimony of somebody who thinks they can speak for them, and the limitation is we have no way of knowing how accurate that is."*

A further risk identified by respondents was that readers might be able to identify themselves or people they knew from the content of posts, and that reading negative descriptions about the impact of autistic behaviors on those around them could lead to guilt or shame. For instance, one participant explained that if they read negative reports from one of their parents on the site, "*that would knock my mental health down the road, because I still feel guilty.*" Uncertainty and uncontrollability about who may be reading posts and what impact those posts could have on their well-being was also raised: *"…my worry is that once someone makes a comment or discusses a situation…there's no control over who [or] how many people consume that content, and what effect it has on them."*

While we noticed a tendency for parents to prefer being able to use the platform to share experiences and autistic participants being more skeptical of this, this was not universal. Instead, the circumstances and purposes of sharing on behalf of others were deemed an important subject importance for further discussions: For example, an autistic participant shared that, *"…thinking about it from that perspective [informing researchers], the idea of my parents writing something about me potentially becomes less disturbing for me."*

https://doi.org/10.1017/dap.2024.21 Published online by Cambridge University Press

Similarly, sharing on behalf of others was not deemed desirable or likely for some parents ("*I don't think I can see me ever doing it [sharing an experience] on his [my son's] behalf*") or was viewed as inappropriate as a principle (*"…it's got to be that person's voice [the autistic person], it can't come from somebody else's opinion, I think that's really important"*). There were also autistic participants who were in favor of parents using the platform: *"I just think people should be able to express themselves and if they want to express what they think someone else experiences in good faith, then they oughtn't feel battered […] into not being able to give that."*

Given this wide range of opinions on the matter, the resolution of this dilemma of whether to allow reporting on behalf of others—and if so under which circumstances—was given a high priority by the larger community and the moderation team.

3.1.4. *Clarity of content submission and moderation guidelines*

Given the complexities outlined above, the need for content submission and moderation guidelines became apparent very quickly, but this came with its own set of challenges. On the one hand, participants felt it to be important that any guidelines would need to have *"a very clear set of rules"* with *"clear criteria"* that would mean that each contribution would just need to be checked for *"does or does it not fall within those criteria."* Such clarity was wished for both to *"make the task of moderation easier"* but also remove *"subjective judgments"* that could frustrate people sharing their contributions.

Beyond the complexity of writing guidelines and rules that would provide such clarity, some contributors also worried about the distinction between intentional rule-breaking, and cases where people sharing their points of view misunderstood the rules and whether sanctions would need to be able to account for such misunderstandings: "*…it literally could just be that the person does not know how to articulate what they were going to say, and it ends up inappropriate, and that just would be discriminatory to remove that post just because they're struggling."* Accounting for the intentions of the person reporting would necessarily make the rules less clear though, as it would require making personal judgments about the writer. For some participants, it was important that decisions were made on a predetermined rule: *"ask the moderators to accept or reject according to whether or not it breaks the rule, rather than subjective judgment.*"

3.2. An iteratively, co-designed moderation approach

As a joint moderation team consisting of autistic and neurotypical contributors, we aimed to resolve these tensions and contradictions as much as possible, making use of solutions suggested by the larger community where possible. Due to their nature as dilemmas, in some cases, different views had to be prioritied as no resolution could work for all different points of views. All moderation guidelines and materials are openly accessible for reuse by other projects. We outline the main features/moderation approaches in response to the identified dilemmas below alongside how they were co-developed with the larger AutSPACEs community.

The main outputs at the end of this process were (1) a set of content submission and moderation guidelines for users of the AutSPACEs platform—specifically tailored to the requirements of our neurodiverse population, (2) analogous guidelines for moderators, and (3) guidance on how to support people in sharing experiences which focus on somebody else while remaining respectful and non-presumptuous.

3.2.1. *No commenting on other users posts*

A high-level decision was made not to allow users to comment on other users' posts, which was implemented both on a technical level (i.e., there is no commenting function) as well as a regulatory level, meaning that the content moderation guidelines specify that creating new posts that are de facto comments are not acceptable. While the analysis of the focus group data showed that some AutSPACEs community members saw potential benefits to allowing commenting—such as creating opportunities for connection and community—the risks that would ensue needed to be balanced against this. In particular, the risks of causing distress or hurt to those posting the original content, either because of negative

https://doi.org/10.1017/dap.2024.21 Published online by Cambridge University Press

content, misunderstanding, or comparative neglect were seen as strong reasons to not allow comments. Furthermore, introducing interactive elements also was seen as carrying the risk of distorting the content, as it would create an incentive to write popular posts, thereby reducing the validity of the data for scientific purposes. Once this dilemma had been uncovered in the thematic analysis, additional community input was collected through the monthly community sessions and discussion on GitHub. In these discussions a consensus for not allowing comments emerged for three main reasons: Firstly, the social pressure to post for "recognition" could bias the scientific data collection. Secondly, commenting would mainly lead to reproducing existing social networks and online fora. And lastly, allowing comments would put an additional burden on content moderators. Given these limitations, the community decided on positioning AutSPACEs as a safe place for all without these pressures which can emerge in more socially interactive online spaces.

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

An initial draft of guidelines for AutSPACEs was presented to all participants in the moderation workshop. These guidelines were written by a researcher based on feedback from previous focus groups and combined elements of two existing codes of conducts: The *Carpentries* code of conduct (The Carpentries, 2019), which was focused on establishing a welcoming and inclusive, open online space; and the AutAngels code of conduct (AutAngel, 2019), which was designed by and for an autistic community. The moderation workshop participants then gave feedback on these draft guidelines and suggested amendments and alternatives. Additionally, a more general discussion took place to better understand the needs, preferences, and concerns of the community.

During the thematic analysis following this workshop, our core content moderation team explored both the direct recommendations as well as emergent implicit needs/requirements. As part of this, we discovered that a member in an early focus group suggested using "tags" for users who wanted to filter out "offensive language" such as swear words without restricting the expressiveness of people submitting their experiences. Given that the analysis also highlighted a broader need for safely sharing and reading, the content moderation co-leads explored such a "filtering" approach in their work, extracting an early list of the kinds of content that focus group participants felt they might find distressing and turning it into filtering categories. The moderation team then presented this early idea to the larger community during the online meet-up sessions, leading to further refinement.

Ultimately, our discussions resulted in a list of categories of potentially distressing content which became the foundation of our "middle-ground" of moderation. Rather than a binary (publish or do not publish), a "traffic light" system to address the dilemma of potentially triggering but at the same time highly relevant experiences was created by the core moderation team. Our goal was to allow users to determine for themselves what sorts of content they would and would not like to see. Our traffic light system breaks down content into three categories—red, amber, and green. Figure 2 gives an overview of the moderation process and how individual contributions are assigned to one of these categories. As part of this process, our guidelines specify which types of content are generally unacceptable, labeling them as "red" and stating that they will not be published on the platform. Examples of such unacceptable content include sharing identifying information (e.g., names, addresses) or actively discriminating against people based on categories such as gender, sexual orientation, or neurodiversity. Content that does not cross the line into unacceptable but that is potentially upsetting or triggering is labeled as "amber." "Amber" content will be published on the AutSPACEs website, but it will include a warning and be hidden by default, allowing users to decide whether or not to view the content of the post on that basis. Examples of "amber" content include content that reports on experiencing discrimination, violence, or abuse, and experiences that share mental health issues or drug abuse. Posts that are otherwise acceptable but contain swear words are also labeled as "amber" content. Lastly, unproblematic content is labeled "green" and will be viewable on the AutSPACEs website by default (see Figure 2). In the current implementation of these guidelines, users can contest moderation decisions which will then be escalated to the core research team.

https://doi.org/10.1017/dap.2024.21 Published online by Cambridge University Press

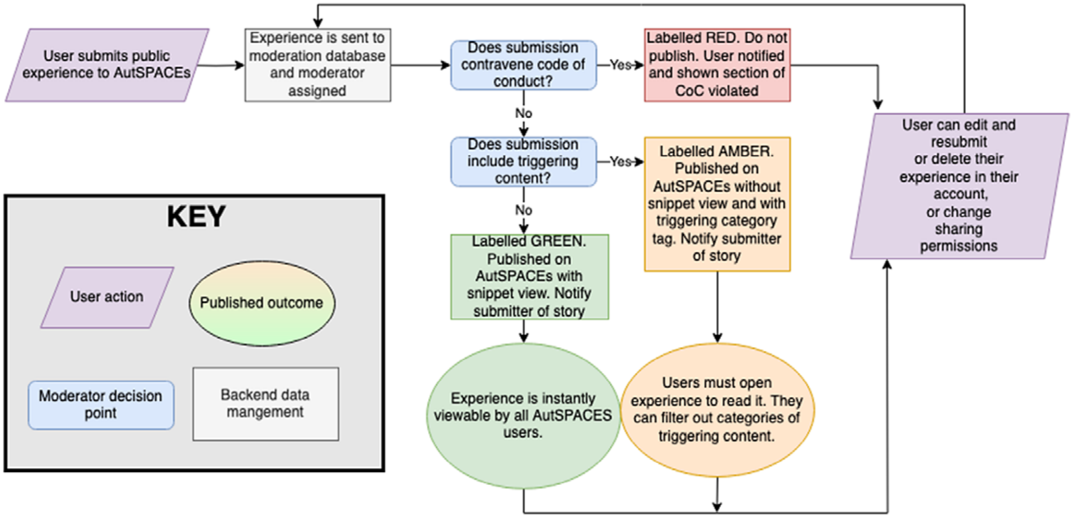
User submits public experience to AutSPACEs
Experience is sent to moderation database and moderator assigned
Does submission contravene code of conduct?
Yes
Labelled RED. Do not publish. User notified and shown section of CoC violated
No
Does submission include triggering content?
Yes
Labelled AMBER. Published on AutSPACEs without snippet view and with triggering category tag. Notify submitter of story
No
Labelled GREEN. Published on AutSPACEs with snippet view. Notify submitter of story
Experience is instantly viewable by all AutSPACES users.
Users must open experience to read it. They can filter out categories of triggering content.
User can edit and resubmit or delete their experience in their account, or change sharing permissions
KEY
User action
Published outcome
Moderator decision point
Backend data mangement

Figure 2. *The content moderation workflow from submitting an experience to publication. Depending on the content, individual experiences are assigned one of three labels: (1) Green, for experiences that are unproblematic. (2) Red, for experiences that are unacceptable. (3) Yellow, for experiences that do not cross into the unacceptable but might be distressing or upsetting.*

3.2.3. Guidelines for sharing stories about others

Following a careful consideration of the multiple perspectives expressed by the AutSPACEs community, it was decided to allow non-autistic people to use the platform to share experiences on behalf of people who are at a particular risk of being excluded, for example, because autism makes it hard for them to communicate directly through a text-based online platform. While an ideal solution would be to adapt the platform so that it could be directly accessed by a more diverse range of users (e.g., with learning differences or who are non-speaking), constraints in time and resources and research needs made such a broader implementation impossible at the current time. Given these limitations, allowing parents and guardians to use the platform to share their observations was considered necessary to include an important cohort of autistic users, even if their representation was indirect rather than direct.

Despite deciding that posts on behalf of others were an overall strength for the autistic community, we were sensitive to the risks presented by allowing non-autistic users to share experiences on the platform. As such, guidelines for supporting others to report their experiences and to share observations about others were produced. These guidelines specify that the focus of the observation should be the autistic person’s experience and how environments might be adapted to improve their experiences, rather than focusing on the non-autistic person who writes the report. These guidelines also specify that the observations should be as neutral as possible, and should avoid making inferences or assumptions about the internal state or motivations of the autistic person being represented. Additionally, non-autistic users are asked to support autistic people to use the platform rather than reporting their own version of the experience and asking for consent wherever possible.

These guidelines for how to support someone in using the platform were the result of extensive discussion in our online community meet-up sessions, departing from a series of “user journeys” that were co-produced to represent some of the different groups of people who may use the platform based on the concerns and desires expressed in early focus groups and the moderation workshop. Based on these community discussions, a set of refined guidelines was produced. In subsequent online meetups, there was substantial consensus that the guidelines were a workable initial solution, leading to them being included as an intrinsic part of the code of conduct and platform design.

https://doi.org/10.1017/dap.2024.21 Published online by Cambridge University Press

To further support non-autistic people using the platform, a series of examples of respectful and neutral versus disrespectful and non-neutral experiences were written in order to help them understand the nuances of sharing on behalf of others. These examples are all based on risks and concerns identified by members of the AutSPACEs community—in particular, concerns about autistic voices being distorted or overtaken by non-autistic people's misinterpretations (i.e., being spoken "about") as well as concerns about negativity about autism and complaints or frustrations with autistic people being common in many online spaces.

This approach of offering guidelines and examples—rather than strict rules—to address the issue of non-autistic users was also the result of the community's input: Realizing the complexity and nuances surrounding the issue, the community members agreed that it would be hard come up with precise rules, given how much of it depends on the user's relationship to the autistic person and the particularity of the circumstances.

3.2.4. When to moderate: pre-publication review

An overall priority, that emerged from the focus groups and moderation workshop, was to allow the respectful sharing of experiences, while making sure users feel safe when reading those experiences. At the same time, the need for clear rules which can account for people misunderstanding them emerged. Given these constraints, the decision was made to implement a pre-publication review/moderation step. This means that all experiences that users want to share publicly are first reviewed by a moderator prior to being published, allowing to assign them following the "traffic light system."

While writing their experiences, users can already report potentially distressing/triggering topics and flag those to the moderators. This is supported by users being shown the moderation guidelines while writing up their experiences. Once submitted for review, an AutSPACEs moderator will review it—using the content moderation to make decisions rather than relying on personal judgment as much as possible. To support this, moderators are shown the content moderation guidelines and examples of types of content which would make a post fall into different categories during the moderation. Moderators also have the option to add additional trigger labels to classify stories as "amber" in cases where the user submitting it might have overlooked appropriate labels. In cases where an experience is flagged as "red" and subsequently not published, moderators will give feedback to the user on which parts of the content guidelines were broken—thus allowing the user to adapt their experience accordingly.

The full moderation flow was presented to the community at a number of online meet-ups, helping to improve both the clarity of the documentation for the users, as well as the workflow for ensuring that contributors whose experiences have been rejected in the moderation process get actionable feedback to re-submit improved/reworded experiences where appropriate.

4. Discussion

Moderating content—in particular for online platforms—is not a trivial task, leading to what has been recently described as the *Moderator's Trilemma* which consists of (1) large and diverse userbases, (2) centralized and top-down moderation policies and practices, and (3) avoiding angering large parts of platform users (Rozenshtein, 2022). In such a setup, it can be impossible to have the "right" degree of moderation, as there will always be some users who feel like speech is being underblocked—while others will feel it is over-blocked, with both views being "right" from the respective user's perspective (Doctorow, 2023).

Rozenshtein (2022) suggests that a federated social networks could be a solution to overcoming the moderators trilemma: Based on the principles of subsidiarity, federation would allow decision-making to happen at the lowest organizational level. Federated moderation systems can be hard to implement, particularly in systems that are potentially of smaller scale or more niche, such as a citizen science project. That is why we focused on bottom-up co-creation with the user community—to design a content moderation policy from the bottom-up rather than implementing it from the top-down—for the AutSPACEs citizen science project. Here, we have described how we used this co-design approach to both collectively identify a number of open moderation dilemmas that are specific to our community in

https://doi.org/10.1017/dap.2024.21 Published online by Cambridge University Press

question as well as—while not per se overcoming them—providing us with a pathway to making informed decisions that are accepted by our diverse, albeit comparatively smaller, user community.

4.1. Co-design to community needs: focusing on safety and support, not engagement

Our resulting approach to content moderation is in many ways quite different from policies and practices found in both larger online platforms as well as other citizen science projects. Increasingly, general-purpose online social media platforms are designed to maximize user engagement with other users and consequently the platform, to the point of leading to addictive behavior (Petrescu and Krishen, 2020), often with the help of algorithms and "dark patterns" (Gray et al., 2018; Waldman, 2020). Similarly, many types of citizen science aim to maximize engagement (Jennett et al., 2016), sometimes using similar "dark citizen science" patterns (Riley and Mason-Wilkes, 2023).

In contrast, the moderation and engagement processes in AutSPACEs look quite different: Instead of encouraging the direct exchange of users by comments through users, such affordances are neither provided on a technical level nor supported through the moderation guidelines. This difference in approach is a reaction to the larger risks that autistic people face around being left out of online interactions and the negative feelings that are associated with this (Triantafyllopoulou et al., 2022). While there would be a potential benefit to receiving supportive comments and reactions, the absence of such support in itself was seen as too big a risk. Instead, collectively we focused on the benefits of providing a safe space to engage in writing in recognition of the fact that the writing in itself can help process challenging situations (Deveney and Lawson, 2022). Similarly, the refusal of supporting real-time publishing of content—as all experiences people wish to share need to pass through the moderation stage—also feed into the additional challenge of bullying that is disproportionally experienced by autistic people (Trundle et al., 2022). With all public experiences undergoing a human moderation approach in addition to a lack of technical support for commenting, the content moderation approach offers two explicit safeguards against bullying. In addition to safety risks to bullying, the co-design of the moderation also identified the need for trigger warning labels, an innovation that is just emerging in content moderation (Morrow et al., 2022), to ensure users can share potentially distressing experiences while keeping readers safe.

Beyond questions of safety, the AutSPACEs community also had to approach the contentious question of representation, in the form of how to support contributions made on behalf of others—if at all. While it has been suggested that the voices of parents and caretakers of autistic people could improve representation (McCoy et al., 2020), this approach has gotten significant pushback: Research shows that autistic self-assessment can differ drastically from assessments done by parents (Hong et al., 2016). As Benjamin et al. (2020) highlight, observations made by parents on behalf of their children are not necessarily a good proxy of their childrens' lived experience, especially compared to the positive impact that the provision of improved communication affordances and resources would have for people that may require them. The moderation approach co-designed for AutSPACEs is aiming to provide such support through guidelines and how-tos that asks non-autistic users to support autistic people in using the platform and sharing experiences where possible, while also providing examples of the nuances of sharing experiences on behalf of others, based on the risks and concerns identified by community members of the AutSPACEs community.

Overall—and in recognition of the importance of embedding autistic views, voices, and values at the core of autism research projects (Botha, 2021)—the non-autistic members of our research team aimed to not make moderation decisions preemptively for the autistic community members, but instead let these decisions be made collectively by the larger AutSPACEs community and its autistic co-creators. As a result of this co-creation process, the content moderation platform for the AutSPACEs platform emerged with a particular focus on creating a safe and welcoming environment.

4.2. Next steps

Based on the existing community feedback, our moderation approach appears to be fit for purpose, but a range of next steps will be performed to put it into socio-technical practice: Firstly, our moderation

https://doi.org/10.1017/dap.2024.21 Published online by Cambridge University Press

approach has now been implemented on a technical level into the first prototype of the AutSPACEs online platform, which will allow us to test this content moderation approach in practice together with the wider AutSPACEs community and explore to what extend this approach is (1) easy to understand and to use and (2) represents the needs and preferences of diverse autistic people, and (3) being consistently interpreted and applied. This work will serve as the basis to iterate our moderation approach in an ongoing dialogue with the community. Secondly, we are also aiming to recruit autistic people as moderators and will develop training and support materials with them, so that they can effectively and safely moderate for AutSPACEs. This will be done with particular attention to the well-being of moderators given the potentially sensitive, upsetting, and offensive submissions they are likely to encounter. Lastly, the moderation for AutSPACEs will remain an ongoing iterative process and all of our moderation documents will be "live" and open to feedback and change throughout the platform lifecycle based on community feedback.

4.3. Adopting a co-design approach

There are a variety of approaches to regulating behavior in online communities and research finds that moderation systems are key to achieve this, as long as the criteria are applied consistently and are accepted by the community in question (Kiesler et al., 2012). We argue that a co-design approach—such as the one we used with AutSPACEs—can provide a clear pathway for designing a moderation system that is in line with community needs. In particular, it allowed us to early on discover particular community priorities that otherwise might have gone unnoticed until late into the implementation phase.

Given our experience, we think that this approach to co-developing approaches for content moderation can be used and translated to a variety of other online communities that share key similarities. A key differentiating criteria for online communities is which niche they occupy in terms of which people they include, which activities they support, and which purpose they have (Resnick et al., 2012). In the case of AutSPACEs, the ambition from the start was to have an online citizen science platform to support a special interest community —rather than being a general-purpose social networking site—for the dual purpose of providing peer-to-peer support amongst autistic people as well as being a research study. In particular, in communities that involve different stakeholders—like researchers and patients in the case of AutSPACEs—such co-creation can be a powerful tool to increase trust (Muller et al., 2021). We suggest that these co-creation strategies that we applied here are particularly suitable for smaller niche communities more broadly.

5. Conclusion

Here, we have demonstrated how we co-created a tailored moderation process for a citizen science project within the AutSPACEs community, moving from community input and data analyses to a set of community governance/policy documents. As part of our process, different key dilemmas, priorities, and issues specific to a diverse autistic population emerged through this extensive, participatory process that included informal discussion sessions and more formal focus groups and thematic analyses. Overall, this participatory approach was a key mechanism for surfacing these complex and nuanced issues that may otherwise have remained unidentified and unaddressed. Understanding and managing these issues was essential to our goal of making AutSPACEs a welcoming and inclusive online space for autistic citizen scientists and our value of empowering autistic people in research.

We strongly believe that our co-creation approach can be adapted and productively be used in other contexts where online platforms and communities are involved, both within other autistic communities but also anyone working on creating platforms for smaller, special interest communities. To support practitioners in implementing this approach, we have made all of our resources and data openly available, and we encourage their re-use by others and to support other collaborative groups.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/dap.2024.21.

Acknowledgments. We thank all members of the wider AutSPACEs community for their support—by giving their time, sharing their knowledge and insights, and contributing to the project across all its different aspects. We thank James Cusack of Autistica for his support and insights that led to AutSPACEs. We thank Sue Iwai and Tom Augur for their frequent input into moderation-related discussions. We are grateful to the developer team of Fujitsu as well as James Kim, Anoushka Ramesh, and Isla Staden for their early contributions to the platform development. We thank Sowmya Rajan, Soledad Li, and Katharina Kloppenborg for their contributions to the design, community engagement, and visual design and Anelda van der Walt for her mentorship support. The team also thanks Amol Prabhu and Lotty Coupat for their project and community management support as civil service fast streamers. We would also like to thank two anonymous reviewers whose input helped to heavily improve the structure and readability of our manuscript. A preprinted version of this article is accessible on SocArXiv at doi:10.31235/osf.io/c2xe7 (Aitkenhead et al., 2023).

Data availability statement. All data and outputs (governance documents, moderation guidelines, quotes of the focus groups, etc.) are available on GitHub at https://github.com/alan-turing-institute/AutSPACEs/. An archived version of the repository at the time of submission can be found on Zenodo at https://doi.org/10.5281/zenodo.10667124. Please see Supplementary Material for a list of all direct links to data from the focus groups, workshops, community meetups, and so on.

Author contribution. Conceptualization: All authors contributed; Data curation: G.A., S.F., J.S.; Formal analysis: G.A., K.W., S.F., J.S.; Funding acquisition: K.W.; Investigation: S.F., J.S., G.A., S.B., B.G.T., K.W.; Methodology: All authors contributed; Project administration: G.A., B.G.T., K.W.; Software: B.G.T., C.M., H.D., D.L.-J., M.S., S.B.; Supervision: G.A., S.B., B.G.T., K.W.; Visualization: G.A., B.G.T.; Writing—original draft: B.G.T., G.A., S.F., J.S.; Writing—review and editing: B.G.T., G.A., S.F., J.S., K.W.

Funding statement. This work was supported by a grant from Autistica to the Alan Turing Institute. This work was also supported by *Towards Turing 2.0* under the EPSRC Grant EP/W037211/1 and The Alan Turing Institute. Fujitsu provided in-kind support for the early platform development. Additionally, Autistica supported the project with in-kind support through co-organizing the focus groups. Otherwise, the funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interest. The authors declare no competing interests.

References

Aitkenhead G, **Fantoni S**, **Scott J**, **Batchelor S**, **Duncan H**, **Llewellyn-Jones D**, **Mole C**, **Smith O**, **Stoffel M**, **Taylor R**, **Whitaker KJ and Greshake Tzovaras B** (2023) How to co-create content moderation policies: The case of the AutSPACEs project. http://doi.org/10.31235/osf.io/c2xe7 (accessed 08 May 2024).

AutAngel (2019) AutAngel Code of Conduct, January 1. Available at https://www.autangel.org.uk/wp-content/uploads/2023/01/230111-AutAngel-Code-of-Conduct.pdf.

Benjamin E, **Ziss BE and George BR** (2020) Representation is never perfect, but are parents even representatives? *The American Journal of Bioethics 20*(4), 51–53. http://doi.org/10.1080/15265161.2020.1730505.

Ben-Sasson A, **Hen L**, **Fluss R**, **Cermak SA**, **Engel-Yeger B and Gal E** (2009) A meta-analysis of sensory modulation symptoms in individuals with autism spectrum disorders. *Journal of Autism and Developmental Disorders 39*(1), 1–11. http://doi.org/10.1007/s10803-008-0593-3.

Botha M (2021) Academic, activist, or advocate? Angry, entangled, and emerging: A critical reflection on autism knowledge production. *Frontiers in Psychology 12*, 727542. https://www.frontiersin.org/articles/10.3389/fpsyg.2021.727542.

Campbell M, **Hwang Y-S**, **Whiteford C**, **Dillon-Wallace J**, **Ashburner J**, **Saggers B**, **Carrington S** (2017) Bullying prevalence in students with autism spectrum disorder. *Australasian Journal of Special Education 41*(2), 101–122. http://doi.org/10.1017/jse.2017.5.

Chambon V, **Farrer C**, **Pacherie E**, **Jacquet PO**, **Leboyer M and Zalla T** (2017) Reduced sensitivity to social priors during action prediction in adults with autism spectrum disorders. *Cognition 160*, 17–26. http://doi.org/10.1016/j.cognition.2016.12.005.

Crane L, **Goddard L and Pring L** (2009) Sensory processing in adults with autism spectrum disorders. *Autism: The International Journal of Research and Practice 13*(3), 215–228. http://doi.org/10.1177/1362361309103794.

Cusack J and Sterry R (2016) Your questions: Shaping future autism research, Autistica. Available at https://www.autistica.org.uk/downloads/files/Autism-Top-10-Your-Priorities-for-Autism-Research.pdf (accessed 08 May 2024).

De Gregorio G (2020) Democratising online content moderation: A constitutional framework. *Computer Law & Security Review 36*, 105374. http://doi.org/10.1016/j.clsr.2019.105374.

Deveney C and Lawson P (2022) Writing your way to well-being: An IPA analysis of the therapeutic effects of creative writing on mental health and the processing of emotional difficulties. *Counselling and Psychotherapy Research 22*(2), 292–300. http://doi.org/10.1002/capr.12435.

Doctorow C (2023) Pluralistic: Solving the moderator's trilemma with Federation (04 Mar 2023) – Pluralistic: Daily links from Cory Doctorow, February 7. Available at https://pluralistic.net/2023/03/04/pick-all-three/. (accessed 08 May 2024)

Feuston JL, **Taylor AS and Piper AM** (2020) Conformity of eating disorders through content moderation. *Proceedings of the ACM on Human-Computer Interaction 4*(CSCW1), 40:1–40:28. http://doi.org/10.1145/3392845.

Fletcher-Watson S and Happé F (2019) *Autism: A New Introduction to Psychological Theory and Current Debate*, vol. *xiii*. New York, NY: Routledge/Taylor & Francis Group, 194. http://doi.org/10.4324/9781315101699.

Gillespie T (2017) *Governance of and by platforms*. In *The SAGE Handbook of Social Media*. London: Sage, pp. 254–278.

Gillespie-Smith K, **Hendry G**, **Anduuru N**, **Laird T and Ballantyne C** (2021) Using social media to be “social”: Perceptions of social media benefits and risk by autistic young people, and parents. *Research in Developmental Disabilities 118*, 104081. http://doi.org/10.1016/j.ridd.2021.104081.

Gray CM, **Kou Y**, **Battles B**, **Hoggatt J and Toombs AL** (2018) The dark (patterns) side of UX design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. New York, NY: Association for Computing Machinery, pp. 1–14. http://doi.org/10.1145/3173574.3174108.

Haimson OL, **Delmonaco D**, **Nie P and Wegner A** (2021) Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation Gray areas. *Proceedings of the ACM on Human-Computer Interaction 5*(CSCW2), 466:1–466:35. http://doi.org/10.1145/3479610.

Haklay M, **Fraisl D**, **Greshake Tzovaras B**, **Hecker S**, **Gold M**, **Hager G**, **Ceccaroni L**, **Kieslinger B**, **Wehn U**, **Woods S**, **Nold C**, **Balazs B**, **Mazzonetto M**, **Rüfenacht S**, **Shanley L**, **Wagenknecht K**, **Motion A**, **Sforzi A**, **Riemenschneider D**, **Dörler D**, **Heigl F**, **Schaefer T**, **Lindner A**, **Weißpflug M**, **Mačiuliene M**, **Vohland K** (2021) Contours of citizen science: A vignette study. *Royal Society Open Science 8*(8), 202108. http://doi.org/10.1098/rsos.202108.

Hong J, **Bishop-Fitzpatrick L**, **Smith LE**, **Greenberg JS and Mailick MR** (2016) Factors associated with subjective quality of life of adults with autism Spectrum disorder: Self-report versus maternal reports. *Journal of Autism and Developmental Disorders 46*(4), 1368–1378. http://doi.org/10.1007/s10803-015-2678-0.

Jennett C, **Kloetzer L**, **Schneider D**, **Iacovides I**, **Cox A**, **Gold M**, **Fuchs B**, **Eveleigh A**, **Mathieu K**, **Ajani Z**, **Talsi Y** (2016) Motivations, learning and creativity in online citizen science. *Journal of Science Communication 15*(3), A05. http://doi.org/10.22323/2.15030205.

Kapenekakis I and Chorianopoulos K (2017) Citizen science for pedestrian cartography: Collection and moderation of walkable routes in cities through mobile gamification. *Human-centric Computing and Information Sciences 7*(1), 10. http://doi.org/10.1186/s13673-017-0090-9.

Kiesler S, **Kraut RF**, **Resnick P and Kittur A** (2012) Regulating behavior in online communities. In *Building Successful Online Communities: Evidence-Based Social Design*. Cambridge, MA: MIT Press. Available at https://direct.mit.edu/books/book/2912/chapter/79067/Regulating-Behavior-in-Online-Communities.

Kloppenborg K, **Ball MP and Greshake Tzovaras B** (2021) Peer production practices: Design strategies in online citizen science platforms, SocArXiv, May 21. http://doi.org/10.31235/osf.io/rw58y.

Landon S (2016) *Romantic Relationships: An Exploration of the Lived Experiences of Young Women who Identify with a Diagnosis of Autism Spectrum Disorder* (pro_doc), University of East London, July. Available at https://doi.org/10.15123/PUB.5531.

McCoy MS, **Liu EY**, **Lutz ASF and Sisti D** (2020) Ethical advocacy across the autism Spectrum: Beyond partial representation. *The American Journal of Bioethics 20*(4), 13–24. http://doi.org/10.1080/15265161.2020.1730482.

Mitchell P, **Sheppard E and Cassidy S** (2021) Autism and the double empathy problem: Implications for development and mental health. *British Journal of Developmental Psychology 39*(1), 1–18. http://doi.org/10.1111/bjdp.12350.

Morell MF, **Cigarini A, and Hidalgo ES** (2021) A framework for assessing the commons qualities of citizen science: Comparative analysis of five digital platforms. SocArXiv. http://doi.org/10.31235/osf.io/pv78g.

Morrison KE, **DeBrabander KM**, **Faso DJ and Sasson NJ** (2019) Variability in first impressions of autistic adults made by neurotypical raters is driven more by characteristics of the rater than by characteristics of autistic adults. *Autism 23*(7), 1817–1829. http://doi.org/10.1177/1362361318824104.

Morrow G, **Swire-Thompson B**, **Polny JM**, **Kopec M and Wihbey JP** (2022) The emerging science of content labeling: Contextualizing social media content moderation. *Journal of the Association for Information Science and Technology 73*(10), 1365–1386. http://doi.org/10.1002/asi.24637.

Muller SHA, **Kalkman S**, **van Thiel GJMW**, **Mostert M and van Delden JJM** (2021) The social licence for data-intensive health research: Towards co-creation, public value and trust. *BMC Medical Ethics 22*(1), 110. http://doi.org/10.1186/s12910-021-00677-5.

Petrescu M and Krishen AS (2020) The dilemma of social media algorithms and analytics. *Journal of Marketing Analytics 8*(4), 187–188. http://doi.org/10.1057/s41270-020-00094-4.

Remmers G, **Greshake Tzovaras B**, **Albert A**, **van Laer J**, **Wildevuur S**, **de Groot M**, **den Broeder L**, **Bonhoure I**, **Magalhaes J**, **Mas Assens S**, **Garcia Torrents E**, **Imre B**, **Covernton E** (2023) Citizen Science for Health: An international survey on its characteristics and enabling factors. http://doi.org/10.31235/osf.io/7tdx5.

Resnick P, **Konstan J**, **Chen Y and Kraut R** (2012) Starting new online communities. In *Building Successful Online Communities: Evidence-Based Social Design*. Cambridge, MA: MIT Press. Available at https://direct.mit.edu/books/book/2912/chapter/79067/Regulating-Behavior-in-Online-Communities.

Riley J and Mason-Wilkes W (2023) Dark citizen science. *Public Understanding of Science 33*, 09636625231203470. http://doi.org/10.1177/09636625231203470.

Rozenshtein AZ (2022) *Moderating the Fediverse: Content Moderation on Distributed Social Media.* SSRN Scholarly Paper, November 23, Rochester, NY. http://doi.org/10.2139/ssrn.4213674.

Salty (2021) Exclusive: An Investigation into Algorithmic Bias in Content Policing on Instagram, October 4. Available at https://saltyworld.net/algorithmicbiasreport-2/ (accessed 08 May 2024).

Sasson NJ, **Faso DJ**, **Nugent J**, **Lovell S**, **Kennedy DP and Grossman RB** (2017) Neurotypical peers are less willing to interact with those with autism based on thin slice judgments. *Scientific Reports 7*(1), 40700. http://doi.org/10.1038/srep40700.

https://doi.org/10.1017/dap.2024.21 Published online by Cambridge University Press

Schaaf RC, **Toth-Cohen S**, **Johnson SL**, **Outten G and Benevides TW** (2011) The everyday routines of families of children with autism: Examining the impact of sensory processing difficulties on the family. *Autism: The International Journal of Research and Practice 15*(3), 373–389. http://doi.org/10.1177/1362361310386505.

Schacher A, **Roger E**, **Williams KJ**, **Stenson MP**, **Sparrow B and Lacey J** (2023) Use-specific considerations for optimising data quality trade-offs in citizen science: Recommendations from a targeted literature review to improve the usability and utility for the calibration and validation of remotely sensed products. *Remote Sensing 15*(5), 1407. http://doi.org/10.3390/rs15051407.

Seering J (2020) Reconsidering self-moderation: The role of research in supporting community-based models for online content moderation. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 107:1–107:28. http://doi.org/10.1145/3415178.

Senabre Hidalgo E, **Ferran-Ferrer N and Perelló J** (2018) Participatory design of citizen science experiments. *Comunicar. Media Education Research Journal 26*(1), 1–10. http://doi.org/10.3916/C54-2018-03.

Skarlatidou A, **Fraisl D**, **Wu Y**, **See L and Haklay M** (2022) Extreme citizen science contributions to the sustainable development goals: Challenges and opportunities for a human-centred design approach. In Ardito C, Lanzilotti R, Malizia A, Larusdottir M, Davide Spano L, Campos J, Hertzum M, Mentler T, Abdelnour Nocera J, Piccolo L, Sauer S, van der Veer G (eds.), *Sense, Feel, Design*. Cham: Springer International Publishing, pp. 20–35. http://doi.org/10.1007/978-3-030-98388-8_3.

Spinuzzi C (2005) The methodology of participatory design. *Technical Communication 52*(2), 163–174.

The Carpentries (2019) The Carpentries Code of Conduct – The Carpentries Handbook documentation, July 17. Available at https://docs.carpentries.org/topic_folders/policies/code-of-conduct.html (accessed 08 May 2024).

The Turing Way Community (2022) The Turing Way: A Handbook for Reproducible, Ethical and Collaborative Research (Version 1.0.2), Zenodo, July 27. http://doi.org/10.5281/zenodo.7625728.

Triantafyllopoulou P, **Clark-Hughes C and Langdon PE** (2022) Social media and cyber-bullying in autistic adults. *Journal of Autism and Developmental Disorders 52*(11), 4966–4974. http://doi.org/10.1007/s10803-021-05361-6.

Trundle G, **Jones KA**, **Ropar D and Egan V** (2022) Prevalence of victimisation in autistic individuals: A systematic review and meta-analysis. *Trauma, Violence & Abuse 24*, 15248380221093689. http://doi.org/10.1177/15248380221093689.

van Schalkwyk GI, **Marin CE**, **Ortiz M**, **Rolison M**, **Qayyum Z**, **McPartland JC**, **Lebowitz ER**, **Volkmar FR**, **Silverman WK** (2017) Social media use, friendship quality, and the moderating role of anxiety in adolescents with autism Spectrum disorder. *Journal of Autism and Developmental Disorders 47*(9), 2805–2813. http://doi.org/10.1007/s10803-017-3201-6.

Veglis A (2014) Moderation techniques for social media content. In Meiselwitz G (ed.), *Social Computing and Social* Media. Cham: Springer International Publishing, pp. 137–148. http://doi.org/10.1007/978-3-319-07632-4_13.

Vohland K, **Land-Zandstra A**, **Ceccaroni L**, **Lemmens R**, **Perelló J**, **Ponti M**, **Samson R**, **Wagenknecht K** (eds.) (2021) *The Science of Citizen Science*. Cham: Springer International Publishing. http://doi.org/10.1007/978-3-030-58278-4.

Waldman AE (2020) Cognitive biases, dark patterns, and the 'privacy paradox'. *Current Opinion in Psychology 31*, 105–109. http://doi.org/10.1016/j.copsyc.2019.08.025.

Wang T, **Garfield M**, **Wisniewski P and Page X** (2020) Benefits and challenges for social media users on the autism spectrum. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*. New York, NY: Association for Computing Machinery, pp. 419–424. http://doi.org/10.1145/3406865.3418322.

Ward DM, **Dill-Shackleford KE and Mazurek MO** (2018) Social media use and happiness in adults with autism Spectrum disorder. *Cyberpsychology, Behavior, and Social Networking 21*(3), 205–209. http://doi.org/10.1089/cyber.2017.0331.

Zalla T, **Barlassina L**, **Buon M and Leboyer M** (2011) Moral judgment in adults with autism spectrum disorders. *Cognition 121*, 115–126. http://doi.org/10.1016/j.cognition.2011.06.004.

Cite this article: Aitkenhead G, Fantoni S, Scott J, Batchelor S, Duncan H, Llewellyn-Jones D, Mole C, Smith O, Stoffel M, Taylor R, Whitaker K and Greshake Tzovaras B (2024). How to co-create content moderation policies: the case of the AutSPACEs project. *Data & Policy*, 6: e28. doi:10.1017/dap.2024.21