“Edit: I’m sorry for being offensive, this is getting downvoted and i feel terrible”: Implicit Social Norms as Governance in Identity-Based Communities

KYLE BEADLE, University College London, UK
MARK WARNER, University College London, UK
MARIE VASEK, University College London, UK

Community norms are important in regulating online identity-based communities as they help shape both online and offline discourse. Yet, not all expressions of identity are treated equally as acceptable forms of speech and expression of identity differ among communities. We used mixed-methods to understand how implicit norms within a set of non-binary communities are reinforced and shaped through influence within these communities. We analyzed approximately 2 million Reddit posts and comments to measure the effect of scores, replies, and self-disclosures, on user editing behaviors, which we use as means to observe norm regulation. We find self-disclosures and the number of replies a post receives is positively associated with editing behaviors, while the influence of scores on the likelihood of a message being edited is highly dependent on whether the message is a post or comment. Our qualitative analysis of posts, comments, and threads finds community norms are created, contested, and reinforced through the interactions between community and individual-level understanding of what it means to be non-binary. We propose a model for implicit norms as governance in identity-based communities, and discuss how platform designers can better use implicit norms to support governance in identity-based communities.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; • **Social and professional topics → Gender**.

Additional Key Words and Phrases: LGBTQ+; online communities; Reddit; social norms

ACM Reference Format:
Kyle Beadle, Mark Warner, and Marie Vasek. 2025. “Edit: I’m sorry for being offensive, this is getting downvoted and i feel terrible”: Implicit Social Norms as Governance in Identity-Based Communities. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW002 (April 2025), 32 pages. https://doi.org/10.1145/3710900

1 Introduction

Online identity-based communities[1] provide spaces for marginalized or vulnerable people to connect and share experiences through self-disclosures [30, 99]. Communities on Facebook, Reddit, Discord, and online forums offer spaces to help individuals disclose their experiences of discrimination and stigma [5, 69, 92] with prior work showing how they support people through acceptance, collaboration, information-seeking, social connection [25, 35, 51, 75, 96]. These communities also enable members to find solidarity and validation among others who share similar challenges and as a result, these online communities often become integral parts of their members’ identities [20, 27, 31].

[1] We define online identity-based communities as informal or formal collectives of people in virtual space joined together by their shared experiences of race, gender, sexuality, nationality, or ethnicity [63, 91].

Authors’ Contact Information: Kyle Beadle, University College London, London, UK, kyle.beadle.22@ucl.ac.uk; Mark Warner, University College London, London, UK, mark.warner@ucl.ac.uk; Marie Vasek, University College London, London, UK, m.vasek@ucl.ac.uk.

ACM 2573-0142/2025/4-ARTCSCW002
https://doi.org/10.1145/3710900

Even a highly public platform such as X/Twitter can enable groups of people to gather around posts about cultural and political issues to engage in activism and community building [61]. These community benefits lead to members changing both online and offline behaviors, reflecting the impact of community attachment, membership, and loyalty on individual users [6, 76, 83, 91].

While they act as important support resources, not all expressions of identity are treated equally within identity-based communities. For example, on Reddit, the r/lgbt community favors posts that create a safe space while the r/ainbow community allows any post to enable “free speech" [41]. Community-based moderation engages members to act as arbiters of what expressions are accepted in the community in addition to enforcing broader platform rules [81]. Moderation practices particularly affect marginalized users and shape community norms around what is acceptable to say in identity-based communities [15]. As a result, moderators in online communities perform significant emotional labor to balance individual expression with individual community nuances [29]. Most prior work on content moderation of identity-based communities has focused on ‘top-down’ content moderation approaches, such as the use of human (e.g., moderators), and technical (e.g., automated content removal) moderation mechanisms [43, 44, 53, 54, 62]. Yet, existing models of social media governance theorize the importance of social norms in the governance of online platforms and their communities. [46, 79]. These models try to capture the diversity of moderation and governance on social media, yet online communities require their own considerations [43, 81], needing moderation to be much more context-dependent as different communities on a platform will hold different values and beliefs, and so have differing preferences toward what content and behavior should, and should not be subject to moderation [59].

We draw on these models to help further our understanding of how informal and implicit community norms are created and reinforced as a means of governance of online identity-based communities, and how platform affordances are used as social influence to help reinforce and respond to these social norms. We also seek to understand the relationship between community norms and self-disclosures as social norms may also be expressed through people editing their self-disclosures after, for example, they misjudge community norms [100] or seek support in sensitive contexts [3]. Where explicit norms are written rules that everyone should adhere to, implicit norms are the informal rules that develop in online communities through daily interactions [13]. Yet, researching social norms of behavior in online communities is challenging due to their implicit nature. Prior work has highlighted how people use the affordances of online platforms to respond to perceived violations of community norms through editing and deleting posts [1, 26, 88, 100, 103]. As such, in this research, we leverage edited posts as a means to understand how implicit community norms are created and reinforced within identity-based communities.

Within this work, we focus on non-binary communities across Reddit. Many online platforms facilitate the formation and maintenance of online LGBTQ+ communities due to their features of pseudonymity and content controls. These features can help protect queer users from being outed, shield them from harmful content, and ease their life transitions [21, 32, 49, 109]. However, these communities also deal with bias within online content moderation systems that can result in unjustified content removal and account deletions [50, 67]. Across Reddit, many LGBTQ+ communities have been established where, for example, LGBTQ+ users can share memes (r/lgbtmemes) and discuss their health (r/transhealth). While gender identity is also discussed in broader LGBTQ+ communities on Reddit, discussions in non-binary communities center on negotiations of different gender expressions–calling attention to the fluidity of the gender norms that we research within this work. Different LGBTQ+ Reddit communities hold different norms of acceptable content [41], and transgender people particularly desire more control over who can and cannot be in their audience [94]. By focusing on non-binary communities, we can include the nuances of non-binary identities into our analysis [19] which allows us to explore how these identity-based communities create and reinforce implicit norms that influence user behaviors.

We conducted statistical analysis on over 2 million Reddit posts and comments to measure the relationship among social influence (up/down votes and number of comments), self-disclosure, and user editing activity. These posts and comments come from ten non-binary communities which we investigate as a case study. Next, we used qualitative methods on a sample of 427 threads, where users provide reasons for their editing actions, to investigate users' motivations behind their editing actions and how users perceive the effects of platform features concerning their edits. Finally, we scrutinize the design implications of our results for platform design, user privacy, and the role of moderators in online communities. Through this analysis, we offer the following three contributions to CSCW:

- We develop insights into how online identity-based communities create and reinforce implicit norms that influence user behaviors.
- We highlight how users of online non-binary communities navigate social media platforms to discuss their identity development.
- We further a framework for understanding how community norms interact with platform affordances as a form of social media governance.

2 Related Work

2.1 Social Norms as Governance in Online Communities

Existing social media governance models theorize the importance of social norms within the governance of social media platforms and the communities that form across these platforms [46, 55, 79]. Grimmelman's social media governance model is based on a "grammar of moderation," where members and moderators are nouns, community characteristics are adjectives, techniques of moderation are verbs, and distinctions of the techniques are adverbs. In this model, norm-setting is one of the techniques of moderation that can be combined with other parts of the grammar of moderation to reach the community's goals [46]. Meanwhile, social norms fit into what Schoenebeck and Blackwell [79] call "normative regulation"–an ecosystem of explicit rules that moderators enforce, and implicit rules that users learn through interacting with the social media platform [79]. They critically analyze the paradigms and harms that occur in social media governance to propose focusing on repairing those harms, such as "reeducation, rehabilitation, and forgiveness" [79]. As such, it is necessary to examine governance within online communities because moderation needs to be context-dependent for each community [43, 59, 81] and online communities of marginalized people are much more vulnerable to moderation harms [67, 94]. For example, transgender people and Black people experienced more content removals and account deletions than other groups [50]. We build upon these models by presenting governance work in online communities to further develop our understanding of the impact of community norms on social media governance.

The governance of online communities relies on the creation and setting of norms, or expected behaviors, for the community to achieve its goals [13]. The study of norms within online communities often focuses on the experiences of moderators and users' reactions to moderation–or explicit norms [41, 54, 66, 82]. For example, Jhaver et al. examine the impact of content removal explanations on user's future activity [54] while Seering et al. investigate how users imitate moderator's behavior [82]. Meanwhile, implicit norms are often studied to examine particular phenomenon [14, 36, 37, 39, 85, 106]. Dym and Fiesler explore how online fandom communities hold norms about ways community members should protect themselves and the larger community [32] and Feuston et al. examine how gender diverse communities hold different norms around discussing eating disorders online [35, 36]. These occurrences of norms highlight how different online communities hold different unspoken values but do not examine the ways these values develop or are enforced. Therefore, we study how implicit community norms are created and reinforced as a means of governance in online communities.

Finally, norms within online communities are measured from either a top-down or bottom-up perspective. Capturing community norms from a top-down approach includes measuring user reactions to explicit rules, content moderation, and demographic changes within the community [16, 38, 54, 60, 64–66, 82, 104]. These relationships are explored statistically, such as Chandrasekharan et al.'s examination of norms through comments removed by moderators [16], and through interviews with community moderators, such as Gilbert's exploration of moderation norms on r/AskHistorians [44]. Even though the approach presented in these two prior works are community-driven, they still rely on a hierarchical relationship between moderators and users. While good at inspecting the impact of community and platform moderation decisions, a top-down approach fails to capture the behavior that emerges from online communities without strict moderation. This approach also fails to distinguish idiosyncrasies among communities that share the same explicit norms. Meanwhile, a bottom-up approach to measuring community norms includes investigating user reactions to other users' posts [14, 27, 32, 37, 39, 85]. This approach is used to understand behavior that emerges from an online community as well as the characteristics of distinct communities. These relationships are explored through interviews, such as Fiesler and Bruckman's investigating norms around copyright in fan-fiction communities [37], and topic-modeling, such as Chancellor et al.'s examination of social support in different weight loss communities [14]. We therefore take a bottom-up approach to examine posts with Reddit communities to conduct an analysis that is grounded in the daily activity of these communities.

2.2 Identity Development, Identity-Based Communities, and Queer Identities

Individual identity, or how an individual defines themselves, is supported by who they relate themselves with and how they navigate those relations. The impact of offline relationships and communities on individual identity has long been studied, with the study of online communities and identity beginning in the 1990s [28]. Since then, online communities have been found to help people develop their identity through discussions that build collaboration [25], support [70], and resilience [30]. The organizational aspects of online communities also impact individual identity [91], such as a user's linguistic style quickly changes after they join a new online community, suggesting that community norms lead to changes in individuals [22]. These online experiences are further affected by an individual's particular identity (such as gender, sexuality, and race) and their current life stage [34]. For example, transgender people in eating disorder spaces are more susceptible to threats to the validity of their eating disorder [35]. Meanwhile, people undergoing life transitions seek separate online communities to navigate how their life transition affects the intersecting aspects of their identity [109].

Managing the presentation of one's identity is particularly important in identity-based communities where people self-disclose more regularly, and seek community acceptance which they do not receive offline [2]. People manage identity and how they present themselves to others through a process defined by Goffman as impression management. This is as the process by which people curate the impression they share with others through intentional behaviors (the impression they "give" and unintentional behaviors (the impression they "give off') [45]. Goffman's work has been extended into the online world with work showing how impression management occurs on different platforms, such as on Facebook [80, 98], Google+ [58], Snapchat [68], and Instagram [5], and how people with different identities undergo impression management through the information people self-disclose, such as young adults [80], veterans [84] and survivors of sexual abuse [3]. Identity-based communities must therefore contend with being spaces that enable vulnerable self-disclosures without creating and enforcing community norms that silence individual expression.

Identity-based online communities provide LGBTQ+ users a safe space where they can explore their identity free from harmful stereotypes and public outings [31]. These spaces allow queer users to avoid the stress of not being supported by people they know offline [48] and aid in the emotional processing of the coming out experience [20]. However, LGBTQ+ people with intersectional identities face difficulties within online communities, as many of these online spaces (such as Reddit) are predominately occupied by white people [4]. Additionally, LGBTQ+ people face homonormativity online from both cisgender, heterosexual people and other LGBTQ+ people [27, 99], such as Taylor and Bruckman who found that bisexual people on Reddit felt excluded and misunderstood by the larger LGBTQ+ community [92]. This finding, where bisexual men were seen as "not gay enough," is the inverse of Devito et al.'s findings of LGBTQ+ being seen as "too gay for Facebook" [27]. LGBTQ+ people must therefore contend their identity with the norms of the online communities they participate in.

Particularly, the experiences of non-binary people in online communities remain understudied in CSCW [93]. Recent work on non-binary people focuses on methodological considerations or including non-binary people within a broader transgender or LGBTQ+ framing [18, 52, 78]. However, non-binary people's experiences are much more nuanced in that they are constantly challenging existing gender norms [19, 97]. Sociological work has found non-binary people situate themselves outside the gender binary by undoing gender and "ungendering" themselves [9, 23]. Conlin et al. found that non-binary people defined their gender experiences in terms of "(a) identity development, (b) heterogeneous identities, (c) identity-expression divide, (d) invisibility and stressors, and (e) resilience and support" [19]. Nova et al. highlight the social media experiences of Hijra in Bangladesh [72, 73]. Meanwhile, Spiel conducted an autoethnography of their own experiences of technological systems which did not allow them to register their gender correctly [89]. These works illuminate how technological systems fail non-binary people, yet non-binary people in online communities work both together and in collaboration with platforms that host those communities.

Importantly, people experience being non-binary in diverse ways. Non-binary people may experience gender dysphoria through their gender/sex, aspects of their body, or not at all [40]. As such, some non-binary people may use the term "transgender" to define themselves while others feel that the term does not fit them [24]. The feeling of a "generational gap" between older and younger transgender and gender nonconforming people further contributes to the nuances of queer identities [86]. Recognizing these nuances is important for community building, such as the preference of Black and Native American transgender and non-binary people to connect with LGBTQ+ communities of color [90]. Therefore, more work is needed to understand the interactions between online community platforms and non-binary people, and how diverse non-binary identities navigate online communities.

2.3 Deletion, editing, and posts that never were

Social media platforms allow for posts and comments to be deleted, with some also allowing the editing of posts. These functions exist to give users control over their online experience and to allow content moderation. Editing and deleting is therefore a retroactive way for users and platforms to manage their accounts, communities, or platforms. However, people may also make substantive edits to what they write before sending, as a means of self-censorship [26, 87].

Prior work has focused on people's editing and deletion behavior concerning their self-presentation and identity management [108]. People edit and delete posts that they consider to be vague, opinions, and sensitive content (such as sex, drugs, and politics) [88, 100, 101, 107]. Sarkar et al. also found that people were more likely to delete a post if they received negative votes from other users [77]. Reasons for editing and deleting are therefore connected to how these topics might affect how other users perceive the poster. Common reasons for editing and deletion include regret, maintaining accuracy, and preserving the feeling of control [88, 100, 101, 107].

People may also edit or delete their message to manage their impression online. People manage their disclosures through edits and deletions to align their online and offline identities [80], to create an ideal identity [33], or to align their identity to the community [58]. Notably, Yılmaz et al. also found that people edit and delete their posts in response to misjudging the social norms of the platform and their connections on those platforms [100, 107]. These findings suggest that implicit community norms may motivate people to more closely manage their impressions online. However, less is known about how these norms are shaped, how platform affordances influence user editing behaviors, and how post-edits maintain these norms. Even less is known about how these behaviors function in identity-based communities where it is known that strong social norms exist [14]. Therefore, a deeper understanding of how these social norms interact with editing behaviors and platform affordances in identity-based communities is needed.

Prior work has also explored the reasons why people edit what they write before sharing a post [26, 87, 103]. People self-censor to manage their self-presentation, avoid arguments, and fit in with their diverse audiences [26, 87]. Privacy concerns also lead people to self-censorship online [103]. Yet while these works find that people self-censor due to the inability to target a specific audience, more work is needed to understand why people may self-censor when they do target a specific audience, such as in an identity-based online community like those found forming around non-binary gender identity.

In summary, while feedback mechanisms that allow communities to exert influence (i.e., social influence) such as up-votes, appear to have an affect on post deletion behavior [77], little is known about how feedback mechanisms affect editing behavior within the context of norm regulation in identity-based communities, and so we ask: **RQ1: How does social influence impact user editing practices in online communities of non-binary users?**

Next, although the relationship between self-disclosure, deletion [100], and self-censorship [26] has been studied, we look to understand the relationship between online norms and self-disclosures, as social norms may be expressed through people editing their self-disclosures [100]. To address this, we ask: **RQ2: How do levels of self-disclosure within a post/comment impact user editing practices in online communities of non-binary users?**

Finally, editing and deletion are all activities where people regulate their social behavior. No prior work, however, explains the implicit community norms that are created and reinforced when people regulate social behavior in identity-based online communities, and so we ask: **RQ3: Why do user's edit their posts/comments and how do editing practices contribute to the creation and reinforcement of community norms in online communities of non-binary users?**

3 Methodology

We investigate the role of social norms and self-disclosure behaviors in regulating behavior through social influence within non-binary Reddit communities. We analyze over 2 million posts and comments across ten Reddit non-binary communities from June 25, 2011, the creation of r/androgyny, until May 31, 2023. We complete our analysis in two parts: a large-scale quantitative analysis followed by a thematic analysis of a sub-sample of our collected messages.

Within our quantitative analysis, we use post editing practices to measure behavior regulation, as prior work shows how people edit posts as a result of affective responses (e.g., feeling regret) [100], maintaining accuracy [88], and preserving the feeling of control [107]. We use post/comment scores to measure social influence, as prior work shows how people are more likely to delete a post if they received negative community feedback [77]. We also investigate how the level of self-disclosure within a post/comment affects the likelihood that a user will edit their post/comment, as prior work shows that people edit their posts to help them manage their impression online [108].

Within our qualitative analysis, we looked to gain a more in-depth understanding into the role of post editing practices in the creation and reinforcement of community norms. We thematically analyze a sample of edited posts, analyzing the reason for their edit, the posts themselves, and the wider threads in which the posts were made.

3.1 Data Collection

To select our included Reddit communities, we consulted an online directory of online, non-binary communities and cross-referenced that list with the most-populated non-binary communities on Reddit that appear through the platform's search function [71]. We used the keywords "queer," "non-binary," "nongender," and "gender-fluid" to identify the communities where queer people shared their experience with gender identity through Reddit's search function. We then used the subreddit's description to determine whether they identified as a subreddit of non-binary users. The amount of communities we included is consistent with other work investigating multiple subreddits [16, 29, 57]. We excluded communities if they were private or they explicitly prohibit data collection in their community rules. An overview of our included non-binary Reddit communities can be found in Table 1

Subreddit	Members	Creation Date
r/nonbinary	214 060	Oct 20, 2012
r/GenderFluid	77 724	May 1, 2012
r/ennnnnnnnnnnnbbbbbby	66 693	Jan 6, 2019
r/NonBinaryTalk	32 108	Apr 6, 2016
r/agender	29 681	Jul 9, 2012
r/androgyny	24 739	Jun 25, 2011
r/NBFashionAdvice	2 459	Apr 11, 2017
r/NonBinaryOver30	1 951	May 23, 2021
r/GenderNonConforming	1 447	Dec 19, 2017
r/AskEnbies	269	Nov 6, 2020

Table 1. Included Nonbinary Reddit Communities as of July 11, 2023

We collected our data from May 8th, 2023 through June 16th, 2023 using Pushshift.io API to gather the submission identifier numbers from our included subreddits and the Reddit API to collect comments [10]. While we seek to use the official Reddit API as often as possible, certain limitations such as rate limit and lack of historical data, required the use of Pushshift.io. The Pushshift.io dataset is peer-reviewed and widely used in published work across disciplines [10]. Baumgartner et al. describe the data collection process as well as their alignment with the FAIR Principles of data management [105]. As of July 2023, Pushshift is no longer available due to changes in Reddit's Terms of Services. However, these changes occurred after our data collection had ended.

From these two data streams, we collected the date, author, title, score, upvote ratio, and body text for all submissions and comments. Before any further data processing, we measure the amount of self-disclosure in each post and comment. Drawing on prior work [8], we use English-language, first-person singular pronouns ("I," "me," "my," "myself," and "mine") as a linguistic parameter to detect self-disclosures. The usage of these terms does not inherently mean that a person is self-disclosing, but does suggest that the user is likely sharing an opinion or personal story rather, for example, providing educational resources or making a joke. Self-disclosure also occurs when users post images of themselves but this behavior is beyond the scope of this study. Additionally, we restrict ourselves to English language subreddits; this naturally excludes non-English language posts and comments from our analysis. We then clean the data by removing posts and comments made by the AutoModerator and other bots, making lowercase all text, removing external URLs, and removing punctuation. The resulting data set includes $n = 342,069$ submissions and $n = 1,823,077$ comments.

The average post across all ten subreddits has 6 comments, 6 self-disclosures, and a Reddit score of 92.39. The average comment across all ten subreddits has 1 disclosure and a score of 5. We recorded 10,306 edited posts or 3.01% of our posts sample while the total amount of edited comments was 55,377, 3.04% of all comments. We present the descriptive statistics of all ten subreddits in Table 2.

			Score per Post			Score per Comment		
Subreddit	**Posts**	**Comments**	**Mean**	**Min**	**Max**	**Mean**	**Min**	**Max**
NonBinary	211 065	1 082 203	88.11 (273.25)	0	5 642	4.86 (17.74)	-435	1 911
GenderFluid	64 070	232 688	42.98 (89.28)	0	1 437	2.04 (2.18)	-25	142
ennnnnnnnnnnnbbbbbby	22 211	233 240	380.70 (550.66)	0	5 119	13.34 (30.42)	-246	1 248
NonBinaryTalk	20 757	123 115	29.67 (53.94)	0	682	6.03 (11.20)	-61	316
agender	18 489	109 544	54.04 (98.84)	0	1 018	4.81 (7.94)	-86	343
androgyny	4 520	38 033	36.81 (52.82)	0	624	2.49 (2.89)	-41	68
NBFashionAdvice	176	744	9.47 (8.48)	0	47	2.47 (1.94)	0	13
NonBinaryOver30	367	1 785	24.27 (24.54)	0	128	2.99 (2.77)	-2	42
GenderNonConforming	347	1 425	12.46 (12.42)	0	80	3.16 (2.81)	-5	24
AskEnbies	67	300	4.45 (2.62)	0	13	2.23 (1.41)	-5	9
Total	342 069	1 823 077						

Table 2. Descriptive Statistics Per Subreddit

3.2 Data Analysis

3.2.1 Regression Analysis. We first perform a regression analysis to answer RQ1 which tests the effect of social influence in online communities of non-binary users. In a Reddit context, scores are the community's reaction to a post or comment—with the score being the number of upvotes minus the number of downvotes [42]. This process results in a post score that everyone can see. Post scores are also presented on a user's profile page next to their posting history. As a result, post scores are more than a public expression of preference toward a post, they are also part of a user's reputation on Reddit. A user with many poorly scored posts will have low Karma, or universal score, and have their future posts removed as many subreddits have Karma requirements to post. Conversely, users with high Karma are viewed more positively. Prior work has shown how users receiving positive feedback in weight loss communities online are more likely to post again in the future [14] while higher post scores in political, online communities result in more positive discourse [74]. We therefore look at the effect of a post or comment's score on the likelihood that a post or comment is edited, hypothesizing that:

H1: A post/comment with a low score is more likely to be edited than a post/comment with high score.

We also investigates the way that comments influence people to edit their post in communities of non-binary users. We only examine the number of comments in relation to posts due to constraints of the Reddit API which did not provide the number of comments for a comment. The number of comments a post receives has been used in prior work as a measure of user participation [64] and to discern the popularity of certain posts [56]. As a result, a post's visibility may increase as the post receives more comments. This may further lead to an unexpected increase in actual or perceived audience size, which Wang et al. found resulted in people regretting their disclosures [100]. To help answer RQ1 we hypothesize that social influence is exerted by a community through reply comments, and that these may result in post edits .

H2: A post with more comments is more likely to be edited than a post with fewer comments.

Finally, we explore how self-disclosures in a post/comment affects the likelihood that a user will engage in self-regulating behavior through post-editing. The effects of self-disclosure within identity-based communities is particularly important because marginalized and vulnerable people develop their sense of self within these communities. Therefore, it is important for people within identity-based communities to be able to self-disclose without fear of retribution or harassment [3]. If someone edits a post where they disclose personal information, they may be regulating privacy boundaries [26, 88] or performing impression management [95]. People may also recognize the archival functions of social media and edit their self-disclosure in order to ensure that their post accurately reflects themselves to future readers [110]. To help address RQ2, we hypothesize that users who self-disclose more in a post/comment are more likely to edit their post/comment.

H3: A post/comment with more self-disclosures is more likely to be edited than a post/comment with fewer self-disclosures.

In all of the tests, we use the variables we present in Table 3.

We use a binary logistic regression to model the probabilities. Logistic regression is the optimal statistical model for this analysis because our dependent variable, whether or not a post or comment is edited, is binary. A logistic regression reveals the effect that a variable, such as a score or self-disclosure has on another variable, whether a post is edited. We run a multiple logistic regression which allows us to assess the the combined effect of multiple independent variables, such as score, on a single dependent variable, such as the editing status of a Reddit message. Furthermore, the inclusion of multiple variables decreases the risk that an unmeasured variable is responsible for the resulting editing behavior. Multiple logistic regression is also transparent–providing various measurements and tests to assess whether the model is fit, whether the outputs are significant, and whether multiple independent variables are correlated.

We take three approaches to our logistic regressions for both posts and comments, found in Table 5 and Table 7. In our first approach, we use our raw data to naively understand the underlying characteristics. A large portion of data received a score of 0, so our second approach (taking the

Variable	Description
ED	Edited status
NC	Number of comments
SC	Message score; the number of a message's upvotes minus the number of downvotes
SD	Total self-disclosure; the number of occurrences of the words ”I," ”me," ”myself," ”mine" within a message
SCB1	Score bucket 11-20
SCB2	Score bucket 21-30
SCB3	Score bucket 30+
SDB1	Total self-disclosure bucket 11-20
SDB2	Total self-disclosure bucket 21-30
SDB3	Total self-disclosure bucket 30+

Table 3. Summary of variables used in regression models

log transform of score and self-disclosure) allows us to better understand the variation within lower-scored posts. Our third approach considers the nuances of different scores and self-disclosure levels at the low end because of the unlikelihood that a post receives a remarkably high score. Here we consider different buckets of scores and self-disclosure ranges to observe these effects.

3.2.2 Qualitative Analysis. Next, we used qualitative methods to answer RQ3 and to further investigate how community interactions and Reddit's platform features influence user editing behaviors in communities of non-binary and gender non-conforming individuals. We used thematic analysis to develop codes for the reasons users provide for editing their message, and perform a reflexive thematic analysis to develop themes from the edit reasons which are made in response to other users [11]. From our initial dataset , we extracted messages (n=4,274) where users explain reasons for their edits, as these provide a ground truth of user's negotiations within a certain community on Reddit. While not an official function, edit explanations arose as etiquette to avoid confusion and to keep users honest about their posting behavior.

3.2.3 Qualitative coding approach. In the first stage of coding, we wanted to understand the broad reason for the edit, and whether the edit was in response to some social influence (e.g., a reply comment), or was standalone (e.g., a user self-correcting a spelling mistake). To do this, an iterative coding approach was used. First, the first author independently generated latent codes on a 10% sub-sample (n=43) and then met with the second author to discuss and refine the the codes. After four rounds, an initial codebook was agreed. A total of 129 edit explanations inform the codebook, and it consists of six codes, three of which are defined as “in response to replies", and three as “standalone'. A detailed codebook can be found in the Appendix. A further sample was then coded by both authors, using the codebook, to ensure reliability of the codebook. The final overall unweighted kappa measured a sufficient 0.87 and so the codebook was deemed to be reliable. As such, the first author independently completed the coding for the remaining 254 edit reasons.

3.2.4 Reflexive Thematic Analysis. We then used a reflexive analysis approach [11, 12] to code posts and comments that were edited “in response to replies” (n=141) as these were made as a result of some external community influence. In all cases, we analyze the edit reason, and analyze additional elements using the below criteria:

(1) If the edit reason is in the original post, we examine the entire thread.
(2) If the edit reason is in a comment with child comments, we examine that comment and its child comments.
(3) If the edit reason is in a comment without child comments, we examine the entire thread.
(4) If the edit reason is in a child comment, we examine the parent comment(s) and the child comments.

We iteratively generated a codebook in three rounds. In the first round, the first author independently generated latent codes, then discussed and refined the codebook with the second author. This was repeated a further two times until no further refinements to the codebook were needed. Finally, the first author used the developed codebook to deductively code the remaining posts, with all codes being reviewed with the second author through discussion, to conceptually check the application of the codebook. As such, no IRR was performed within this analysis. Finally, the first and second authors then iteratively group the codes into broader themes until they are richly defined.

3.3 Ethical Considerations

We received departmental ethical approval for this project. Additionally, our project seeks to understand the impact of different platform features on Reddit, and we seek to maximize the benefit of this project by highlighting the effects that social networking site platform features have on individual privacy, censorship, and self-disclosure in online communities [7]. By consenting to post on Reddit, users consent to Reddit's Terms of Service and agree that their posts, comments, and more may be accessed through the Reddit API. Our data collection occurs much later than when users post to Reddit. Therefore, we assume that all the data collected is moderated and follows Reddit's Terms of Service.

Non-binary people use Reddit as a safe space to explore their gender identity because it is not always safe to do so offline. By collecting textual data on these discussions, we identify risk in “outing" users [47]. The collected data is pseudonymous and any data that links a pseudonymous user with an identifiable person is deleted.

We paraphrase all quotations we present here to reinforce user anonymity. All quotations were paraphrased by the first author, and the second author ensured that the paraphrases maintained the messages' original meaning.

3.4 Limitations

We recognize potential limitations in the specific measurements used, the demographics of Reddit, and the sampling method for our qualitative analysis. We chose to quantitatively measure the effect of comments, scores, and self-disclosures on the likelihood that a specific Reddit user would edit their post. While our measurements focus on the *general* normative behavior in online communities of non-binary people, future work examining the effect of other measures, such as toxicity, could illuminate how specific norms are developed through helpful or hateful language. Because of the diversity of non-binary identities present in our data, our findings are able to speak to Reddit as a technology which supports non-binary people as well as the challenges of governing identity-based communities where people come from heterogeneous backgrounds. We note that Reddit racial demographics skew white which may be reflected in our data. Future work could extend our study to understand how implicit norms operate in other marginalized communities such as those supporting racial minorities. Furthermore, our qualitative work considers Reddit threads which include a post/comment where the poster explains why they edited their post/comment. This approach excludes the many unspoken reasons people may edit on Reddit. Despite these limitations, examining edit explanations enabled us to analyze the dialogue that occurs between a user and a community.

4 Quantitative Results

To address our first research question, we first examine our results from Reddit posts and comments to determine the effect of scores on the likelihood of a post or comment being edited (H1). Then, we examine our results on the effect of reply comments on the likelihood of a post being edited (H2). To address our second research question, we examine our results on the effect of self-disclosure on the likelihood of a post or comment being edited from (H3). In this section, we first present the findings on editing influences on posts, and then present findings on edit influence on comments (i.e., replies to posts).

4.1 Analyzing effects on editing behavior on initial posts using regression analysis

We want to understand the effect of scores, number of comments, and level of self-disclosure in a post, on the likelihood that a user edits their post. We use three separate logistic regressions to address this; this shows the robustness of our approach and allows us to uncover lurking trends in our data as our variables are not correlated (Table 4. Figure 1 shows our general approach, and our variables are defined in Table 3.

	ED	**SC**	**SD**	**NC**
ED	1			
SC	-0.046	1		
SD	0.095	-0.066	1	
NC	0.025	0.619	-0.011	1

Table 4. Correlation table for our variables on Reddit posts.

$$Pr(\text{ED} = \text{True}) = \frac{\exp(\beta_0 + \beta_1\text{SC} + \beta_2\text{SD} + \beta_3\text{NC})}{1 + \exp(\beta_0 + \beta_1\text{SC} + \beta_2\text{SD} + \beta_3\text{NC})}$$

Fig. 1. Logistic Regression for Posts (Regression (1) from Table 5).

Our results (Table 5) show that there is a negative relationship between a user's editing behavior (ED) and the score of any given post (SC). This finding is robust against our three models. This means that users with low post scores are more likely to edit their posts than users with high post scores. Specifically, for each additional upvote (a single increase in score), the likelihood of the post being edited *decreases* by 0.99 ($e^{-0.0142936}$) (Table 5 (1); $p < 0.001$)[2].

We also find that there is a positive relationship between the number of comments a post receives (NC) and the likelihood that the post will be edited (ED). In other words, the amount of replies on a post influences user editing behavior. The likelihood of a post being edited *increases* by 1.05 ($e^{0.0487956}$) for each additional comment to a post (Table 5 (1); $p < 0.001$).

[2] We calculate this by inserting our calculated β into the equation in Fig. 1.

Dependent Variable: Edited Status of Post (**ED**)

	(1)		(2)		(3)
NC	0.049***	**log(NC)**	0.87***		
	(0.00098)		(0.012)		
SC	-0.014***	**log(SC)**	-0.57***	**SCB1**	0.054***
	(0.00035)		(0.0085)		(0.029)
				SCB2	-0.33***
					(0.044)
				SCB3	-1.1***
					(0.030)
SD	0.015 ***	**log(SD)**	0.31***	**SDB1**	0.812***
	(0.00041)		(0.0079)		(0.030)
				SDB2	0.97***
					(0.037)
				SDB3	1.22***
					(0.030)
$R^2 = 0.069$		$R^2 = 0.099$		$R^2 = 0.049$	

*** Results significant at the $p < 0.01$ significance level.

Table 5. Results for our three binary regression models on posts. Standard errors in parentheses.

Meanwhile, we show that a user's editing behavior is positively associated with the amount of self-disclosure (SD) in a post. The likelihood of a post being edited *increases* by 1.02 ($e^{0.0153095}$) for each additional self-disclosure in a post (Table 5 (1); $p < 0.001$).

These results support our hypothesis (H1) that a low-scoring post is more likely to be edited than a post with a high score. The subsequent regressions in the table show the robustness of the result is; particularly regression (3) shows that it is robust to approximate rank of the score, rather than the actual number itself. As such, we add evidence towards our hypothesis that online communities use downvotes/upvotes to influence a user to edit their post, thereby reinforcing community norms. For instance, a user posted that they don't feel guilty when they call people out for misgendering them in a post entitled, *"Do you think I am a bad person?"* They received downvotes (resulting in a low post score) and edited their post to address them, *"Edit: I don't know why I'm being downvoted, but I will correct anything if someone explains why I am wrong here."* While we see that posts with low scores are more likely than posts with high scores to be edited, these results are unable to show us how or why low scores influence user's editing behavior. For example, posts may have low scores because of downvotes or a general lack of engagement with a post. It is also unknown how users may respond to either of these outcomes. We therefore qualitatively investigate why and how users respond to their posts receiving low scores, where they have made an edit, in Section 5.1.2 and Section 5.3.1.

Next, we see that posts with a high number of comments are more likely to be edited than posts with a low number of comments, supporting our hypothesis (H2). For example, on a post with 45 comments, the poster edited their post to say, *"Edit: Finding the term 'agender' changed my life. Thank you all for the replies! I needed all your validation right now."* This adds evidence towards our hypothesis that comments are also used in online communities to reinforce community norms. A post with many comments signals a high level of community engagement with a user's post, but it is unclear what type of comments influence a user to edit their post. Comments may be overwhelmingly positive, negative, or be more mixed in tone. Similarly, a single negative comment could affect a user more than many positive ones. To further investigate the role of reply comments in creating and reinforcing community norms, we explore community comments qualitatively in Section 5.

Finally, these results support part of H3 that a post with more self-disclosure is more likely to be edited. This result suggests that community norms exist around the level of self-disclosure a user includes in their post. For example, a post entitled, *"I'm embarrassed to tell people my chosen name"* included 47 disclosures. The community responded by asking for more information about the user's experience and they edited their post to answer, *"Edit: I have anxiety, dysphoria, and am agender. Sometimes, I wish I didn't have a name at all."* However, less is understood as to why posts with more self-disclosures are more likely than posts with low self-disclosures to be edited, something we explore qualitatively in Section 5.1.3.

4.2 Analyzing effects on editing behavior of subsequent comments using regression analysis

Similarly to our previous subsection on edited original posts, we want to understand whether scores and self-disclosures are used to influence users editing behavior via *comments*, addressing RQ1 and RQ2 as well as our hypotheses H1 and H3. As before, we use three separate logistic regressions to address this; this shows the robustness of our approach, and allows us to uncover lurking trends in our data as our variables are not correlated (Table 6). Figure 2 shows our general approach and our variables are defined in Table 3.

	ED	**SC**	**SD**
ED	1		
SC	0.028	1	
SD	0.131	0.003	1

Table 6. Correlation table for our variables regarding Reddit comments.

$$Pr(\text{ED} = \text{True}) = \frac{\exp(\beta_0 + \beta_1\text{SC} + \beta_2\text{SD})}{1 + \exp(\beta_0 + \beta_1\text{SC} + \beta_2\text{SD})}$$

Fig. 2. Logistic Regression for Comments (Regression (1) from Table 7).

Our results in Table 7 show that a user's editing behavior is positively associated with the score a comment receives. In other words, the amount of upvotes a comment receives and the amount of self-disclosures influence a user's behavior in editing their comment. This finding is robust against our three regressions. Specifically, for each additional upvote (a single increase in score), the likelihood of the comment being edited *increases* by 1.00 ($e^{0.0046139}$) (Table 7 (1); $p < 0.001$).

We also find that editing behavior has a positive association with the amount of self-disclosure in a comment. Particularly, we find that the likelihood of a comment being edited *increases* by 1.10 ($e^{0.0934138}$) for each additional self-disclosure in a comment (Table 7 (1); $p < 0.001$).

Unlike post edits that were negatively influenced by score, reply comments were positively influenced by score which does not support part of H1. Therefore, comments are more likely to be edited with each additional upvote received. For example, a comment where a user shared their personal experience of being non-binary in Chile received a score of 193, and was edited to say, *"All of you nonbinary folks made my day! Thank you!."* This result suggests that different community

Dependent Variable: Edited Status of Comment (**ED**)					
(1)		(2)		(3)	
SC	0.0046^{***}	**log(SC)**	3.4^{***}	**SCB1**	0.67^{***}
	(0.00015)		(0.093)		(0.021)
				SCB2	0.82^{***}
					(0.033)
				SCB3	0.96^{***}
					(0.025)
SD	0.093^{***}	**log(SD)**	0.86^{***}	**SDB1**	1.8^{***}
	(0.00078)		(0.0060)		(0.022)
				SDB2	2.4^{***}
					(0.039)
				SDB3	2.8^{***}
					(0.046)
$R^2 = 0.040$		$R^2 = 0.061$		$R^2 = 0.036$	

*** Results significant at the $p < 0.01$ significance level.

Table 7. Results for our three binary regression models on posts. Standard errors in parentheses.

norms exist for comments than for posts. To pull apart the differences between the effect of post scores and comment scores on editing behaviors, we qualitatively investigate user reactions to downvotes in Section 5.1.2 and Section 5.3.1.

These results also support H3 as they show how comments with more self-disclosure is more likely to be edited. For example, a user responded to the post, *"What does it feel like to be nonbinary,"* with 22 self-disclosures and edited their comment to say, *"Edit: It feels very personal to me. I am gender-fluid and don't care about pronouns. It feels like freedom."* We see the same effect present in our previous result where Reddit posts with more self-disclosures are more likely to be edited than posts with fewer self-disclosures. This suggests that the communities have similar norms for self-disclosures in comments as they do in posts. We further examine community reactions to self-disclosures in Section 5.1.3.

5 Qualitative Findings

While we quantitatively find that user behaviors, such as upvotes/downvotes and number of comments, influence a user's likelihood to edit their post or comment on Reddit, we also want to understand the qualitative experiences of users who have edited their posts.

Through our analysis of users' optional in-text explanations of their editing behavior, we identify strategies in which implicit community norms (not explicitly written down as Community Rules) guide acceptable behavior in subreddit communities of non-binary users. We find that community norms exist around the behavior and language embedded in posts that the communities deem acceptable. The boundaries of these norms are then the lines that the communities draw between acceptable and unacceptable behavior and language. We find that community norms are created, contested, and reinforced through the interactions between community-level and individual-level understandings of what it means to be non-binary. As such, we detail two themes that influence user editing behavior: community consensus and community uncertainty. The theme of community consensus describes how non-binary communities *reinforce* the general agreement, or norms, of the group. Meanwhile, the community uncertainty theme describes how non-binary communities manage disputes and *create* general agreement, or norms, through group discussions.

Finally, we find that some platform mechanisms and discussion strategies influence users regardless of whether they face community consensus or community uncertainty. We detail user reactions to the these strategies and mechanisms and, similar to the influential user behaviors we uncover in Section 4, we describe how online communities use these strategies and mechanisms to create and reinforce community norms.

5.1 Community Consensus

The first theme that we developed relates to responses to edited posts that show a level of community consensus that has developed around content and language use. When a user creates a post that shares their understanding of what it means to be non-binary, the community often judges whether or not a post fits within its acceptable community norm boundaries. We see users editing their posts in response to the community rejecting, accepting, and being impartial toward their individual understanding of what it means to be non-binary, embedded within their posts. For each sub-theme, we first describe the types of posts where community consensus exists and then describe user responses to community consensus, observed through editing behaviors.

5.1.1 Community rejection of users or posts. In the edited posts analyzed, we found some to be embedded within threads that included responses from the community that were consistent in their *rejection* of the post's content, suggesting a level of community consensus had developed around the topic being discussed within the edited post. For example, after a user creates a post, *"Men and women have different brains,"* the community responded with significant downvotes and comments such as, *"I hate this comment,"* and *"This just isn't true and is used in TERF and NB-erasing arguments."*

Rejection is used by communities to help reinforce community inclusivity and address misperceptions of what it means to be non-binary. For example, when a user spoke about gender expression not being a choice, another user corrected them by saying, *"gender expression IS a choice, but gender identity is not. The way someone chooses to dress or do their hair is a choice. What someone is (their gender identity) is not a choice. Trans men may express as femme and still be a valid man."* Such errors made in good faith may lead to the community rejecting the post, but not the user who created it. In some cases, we found the community responding to users in an attempt to correct their views. This community behavior reinforces definitions of gender identity and expression that the communities hold at large.

In some cases, rejection occurred when the personal experience of a user or group of users contradicted the opinion expressed within the edit post. When one user posted their confusion as to why non-binary people may want to medically transition. They were opposed by another user who shared their personal experience, *"Even though I'm non-binary, HRT will allow me to have less body hair and more fat on my hips. My presentation is not 'binary' because of this."* Here, we find individual opinions being challenged by the community through the use of personal experiences, which reinforces the importance of consistent community participation.

User posts were not only rejected due to the views expressed within them, but also the language that the community considers to be unacceptable. When one user, new to the community, felt excluded from women's spaces because they were not *"socialized as a woman,"* community members were quick to inform that they should not use the phrase because *"TERFs tell transwomen that they are 'socialized male' and not real women which is not true."* In this case, community rejection is used as a safety mechanism against bad actors and individuals wanting to harm non-binary individuals. The excluded user later apologized for their comment and stated that they did not know better, as seen in Section 5.1.2; the community eventually accepted the user. However, this highlights the tension between the importance and opacity of implicit community norms as these norms exist to determine genuine community members from bad actors such as trolls and transphobic people, yet their opaque nature can lead to confusion where genuine newcomers, not yet familiar with these norms, are rejected by the community they seek to engage with. These findings help us to answer RQ3 by highlighting that users edit their posts toward community norms in community discussions of inclusivity, opinion-based discussions, and discussions that fail to use community-approved language.

5.1.2 User responses to community rejection. Above, we describe threads where there was consensus around rejecting posts due to their content, or the language used. Where that rejection occurred, we found users responding to community-level rejection by editing their posts to adhere to community norms. For example, the below edit reason describes an edit made to conform to community norms around acceptable language: *"Edit: I've been informed that the phrase "socialized as a woman" is terf rhetoric. going forward, I'll avoid using it. I never, ever want to come across as marginalizing trans women. I appreciate all of the responses."* Rejected users also edited their posts to reflect new evidence presented by other users. Users particularly responded to medical evidence to inform their editing behavior, such as when one user posted about the potential negative impacts of a hormone replacement drug, bicalutamide. After receiving more information from the community, the original user responded with a correction: *"Edit: confused bicalutamide with other classes of drugs."*

After rejection, users editing their posts may be seeking acceptance from the communities or may signal that they are amenable to community norms, as seen in the edit reason above. Editing a post in line with community norms shows the community that a user is acting in good faith and willing to conform to the implicit community norms developed within the community. For example, one user deleted their post after a community rejection and commented, *"Edit: Apologies for the misunderstanding, I gladly delete this post. Edit 2: In the future, I will post more positive things."*

These findings help us to answer RQ3 by illuminating the prevalence of community rejection toward users as one of the reasons that users edit their posts on Reddit. As such, strong norms exist in non-binary communities around acceptable language, deference to medical literature, and preference for personal experience over opinion.

5.1.3 Community acceptance of users or posts. In contrast to threads rejecting posts, we found threads where the community was consistent in their *acceptance* of posts. We find non-binary communities on Reddit signaling their acceptance of posts through upvotes and supportive comments. For example, when a user asked if they, as someone who is cisgender and straight, could use they/them pronouns, one community member responded, *"Do it! No matter your identity, gender expression and pronouns are meant to be played with!"* When the communities accept a user's post, they reinforce community norms associated with the post, and in doing so reinforce that community as a safe space and resource for non-binary users.

The communities accepted coming out stories, personal questions, and selfies seeking affirmation. One user shared their experience of quitting their job after coming out and their boss refusing to use their correct pronouns. The community responded to this story, with one user writing, *"What a badass! You will certainly find a better job. Keep standing up for yourself and seeking happiness. Everything else will fall into place."* Additionally, if a user posted a personal question, such as, *"Am I non-binary or am I just pretending,"* community members often answer by providing their own personal stories. For example, in this case, a user responded, writing: *'Gender is like shoes for me. The "girl" shoe just doesn't fit. Once I found my non-binary shoes, I finally felt like I had more room in my gender.'* Finally, when commenting on selfies, community members pointed out aspects of the photo they found pleasing. One user commented on such a post saying, *"You are valid as fuck, and that color palette looks great on you!"*

In Section 5.1.1, we see communities rejecting opinion-based posts in favor of posts that share personal experiences. We see the inverse of this phenomenon in community acceptance. Here, communities do not just *favor* personal experience, they strongly *accept* it. These communities were founded for inter-personal support, so it is unsurprising to see the communities accept posts with personal experience. Newcomers, who are uncertain about their identity, may feel uncomfortable with this practice, though. As such, sharing personal information may favor long-term community members and users who are in safer offline situations. Discussions around coming out stories, personal questions, and selfies seeking affirmation are therefore edited to respond to norms of community acceptance, helping to answer RQ3.

5.1.4 User responses to community acceptance. We find users appropriating the practice of editing and writing edit reasons into their posts to signal their gratitude to the community and to share more information with the community. For example, here we see this user responding to the community, thanking them for their support, writing: *"Edit: I'm still discovering myself, so thank you everyone! I appreciate your help and kind words."* Once accepted, a user may feel overwhelmed by the supportive messages they receive online that they do not receive offline, such as one user who wrote: *"Edit: These responses are overwhelming. You all made my night with your creativity and awards. Thank you!"* Users editing their posts to respond to acceptance also gave back to the community. In response to the community wanting to create more social connections with other non-binary people, a user edits their post to include information about a new communication channel that they had set up, writing: *"Edit: I created a Discord server because it appears that many individuals are also hunting for enby pals! Get a link by messaging me!"* These responses may also reify community connections and obscure community knowledge from newcomers who were not in the community at the time. These findings help to answer RQ3 by clarifying that community norms of acceptance toward users are one of the reasons that users edit their posts on Reddit.

5.1.5 Community impartiality towards users or posts. We found users editing behaviors were not only influenced by positive and negative community responses but also by impartial community responses. Community impartiality refers to the objective treatment of a user/post within an inclusive discussion. These posts occur within existing community consensus, and, therefore, maintain it. Impartially received posts were often edited to add more information to the original post, typically in direct response to requests received from the community within their replies. When a user asked for suggestions on a new name for themselves, other users wanted more details about what they were looking for in a new name, asking, *"Do you want a more feminine name or something more neutral? I picked a more more natural name for myself, so maybe you want something similar?"* As we have seen around positive and negative responses, here we again find the community responding by drawing on their personal experiences to better understand the position of the original poster. Similarly, when someone asked about how they could pass as a non-binary person without hormone replacement therapy, another user replied, *"Why are you so concerned with passing? Is it for your own safety? Is it so you don't get socially rejected? You should consider all of these things."* Similarly to how community rejection was used as a safety mechanism, here we see impartial responses being used to ask questions of the original poster to better understand the safety implications around the post. This finding helps answer RQ3 by calling attention to the importance of community safety norms causing users to edit their posts.

5.1.6 User responses to community impartiality. We find community impartiality connected to the maintenance of community norms as users learn more about one another to better understand the context of their discussions and its safety implications. In response, we find users editing their posts to respond to these community requests for more information. For example, one user stated that they were strongly against medical transitions and wanted advice on managing gender dysphoria. After another user voiced their support for this initial poster, the second user returned to give even more support after the initial poster shared more doubts: *"Medical transition can be great because it gives you autonomy. You are creating your own body by choice. My body is now my own because of how I want it, not how society views it. I 100% support people who don't medically transition. You need to think about what makes you dysphoric and working on what will alleviate that."* While the communities neither strongly accept nor reject these two posts, they may offer the original author more freedom to discuss their issue before the communities decide on the post. This impartiality may take the form of hesitancy which stems from the community's concerns about safety. After a user told the community that they did not like how they were treated as a woman, they asked if they could be non-binary instead. One user commented, *"You may need to read up more about what it means to be non-binary. It's about feeling neither male or female, not how society treats women."* While neither rejection or acceptance, the user who replied edited their comment to ask, *"Edit: Do you know that people think non-binary people are liars,"* questioning the intentions of the original poster. Therefore, as the user edits shown above demonstrate, users try to indicate their amiability to the community when responding to community impartiality. These findings further help to answer RQ3 by explaining how the community may influence users through norms to edit their posts when neither strong acceptance nor rejection is present.

5.2 Community Uncertainty

Where community norms are not yet well formed, community discussions are used to manage disputes and create new social norm boundaries that the community may adopt. In these cases, uncertainty arises in the differences in the way the communities should respond to a post. Posts that lead to such uncertainty contain a split between users who want to reject or accept a user for their post. We find community uncertainty exists largely around discussions of community exclusivity, terminology, and physical expressions of being non-binary.

Discussions of community exclusivity arise when users notice trends in the community that they believe violate community inclusivity. For example, when one user pointed out the communities' focus on each others' appearances, *"This subreddit is obsessed with androgyny and aesthetics. That's not what being non-binary is about,"* replies were then split between users agreeing with the post and users addressing the importance of aesthetics. The terminology and labels used by community members were also highly contentious, with one user asking, *"Do popular non-binary terms bother anyone else?"* Users were again split between whether they prefer terms such as *"datemate"* to other terms such as *"partner."* Finally, the communities often debated how to physically express being non-binary and how the communities valued certain expressions. Users felt that they appeared too masculine, *"Non-binary people assigned male at birth ARE non-binary people. End of discussion,"* or too feminine to be non-binary, *"Do all cis women hate their sex characteristics as much as I do?"* Users in these discussions were split between whether or not there was a larger gatekeeping problem in the non-binary people–with some users feeling that non-binary people assigned female at birth were more welcome than those assigned male at birth.

These themes and their corresponding responses answer RQ3 by revealing the 'wrangling' a community undergoes to gain some consensus over acceptable content and language. Community uncertainty leads to users justifying their stance, maintaining their stance, or changing their stance.

5.2.1 The user justifying their stance. In some cases, we find users who are faced with community uncertainty acting to edit their posts to justify their stance. Justification, in this context, occurs when users seek to motivate their stance–presenting the underlying assumptions, emotions, and experiences that they hold. Users present their underlying assumptions when there is community uncertainty around their intentions. For example, in a post that included a question related to pronouns, some users dismissed the author of the post as being ill-informed, while others embraced the chance to educate a new community member. Addressing this discussion, one user edited, *"Edit: To clarify, while it's okay to ask questions, it's best to do a little research before so that you can come up with an actual question rather than merely asking us to justify our existence."* Users edit their posts with emotional justifications to convey the importance of a particular topic. After a user deleted their post about gatekeeping in the community, another user edited their post to justify their frustration at the lack of open discourse, *"Edit: All they wanted was to draw attention to a problem that needs to be addressed in this community. It makes me sad to see that the post was deleted."* Finally, users draw upon their personal experiences to present stronger evidence for their initial argument. One user edited their post to share that the people around them offline support their position, *"Edit: The people around me are confused if I tell them that I am trans. They think that I identify as male even though my discomfort comes from being either gender."* Revealing underlying motivations, conveying emotional impact, and sharing personal experience suggests that users engage with community uncertainty to convince other users of their position, helping to answer RQ3. Therefore, individual justifications are involved in community norm-setting as users attempt to form a consensus around their stance.

5.2.2 The user maintaining their stance. Finally, in the face of community uncertainty, we find users maintaining their stance through their edits. Where these justifications provide explanations for a user's stance, we find users maintaining their position within post-edit descriptions, through the reiteration of their original post, and continued discussion. A user that reiterates their stance may comment on the whole discussion as it relates to their perspective, such as the user who, when arguing that non-binary people should not use the letters *"nb"* because they also stand for *"non-black"* reiterated their point by editing, *"It's disappointing to see so many queer appropriators."* Such a comment is both a reiteration of their original stance, as well as a comment highlighting their frustration at the lack of consensus within the community around this particular topic. Finally, users may maintain their stance by continuing to push a discussion forward. One user received mixed reactions from the communities after posting a transphobic comment they saw on another social media platform; they edited their post to say, *"Edit: No, after giving it some thought. This conversation is necessary because we must be able to recognize potential reactions in others and discuss them. I do want to have a serious conversation about how to handle negative comments. Sorry if you disagree, but I believe that in order for people to be able to address it when it directly affects them, they need to be aware of this."* While the community may be uncertain in their response to certain topics and language, continuing to push a discussion may compel a community to create new settled norms. As such, users also edit their posts to show their persistence in trying to convert the communities to accept their posts, furthering RQ3.

5.2.3 The user changing their stance. We also found users changing their stance as a result of persuasion from other users, or where new evidence challenged their initial view. For example, in a discussion about pronouns, a user returned to their post, adding an edit to state, *"Edit: I think you did a great job explaining this after giving it some thought last night and this morning. I believe I just need to work harder on this, as I haven't really given it enough attention, believing my own gender identification to be a kind of "pass" on my prejudices."* In another example, a community discussion occurred around a photograph that presented the symbols of 12 different gender identities. In response, one user posted a statement stating that *"Most of these genders sound fake"*, later editing their post to retract their statement having conducted wider research on the topic. In this case, the user's edit wrote: *"Edit: I retract my statements after doing some looking."* This finding helps to further address RQ3 by explaining that users changing their stance as a process of community norms becoming solidified.

5.3 Mechanisms for Social Norms Regulation

Across the themes of community consensus and community uncertainty, we find mentions of specific platform features (e.g., downvotes) and discussion topics (education and awareness and support) that influence whether a post is edited. We find different mechanisms being used to support communities in creating and reinforcing their social norms.

5.3.1 Downvotes. We see Reddit's downvote feature being used in community rejection of the user or post, where users justify their stance, and where individuals act to maintain their stance. These findings help to answer RQ3 by illustrating the various ways in which downvotes, as a platform feature, are used to enforce community norms and influence users to edit their posts.

When a community rejects a post using downvotes, users respond by changing the content of their post. When one user received downvotes for a post dismissing teenagers who *"make up genders,"* the same user changed their post with the edit, *"Edit: Let me explain because this sounds more inflammatory than I intended."* Therefore, communities use downvotes to signal community rejection and further reinforce norms of acceptable language.

Downvotes also result in users justifying their stance. Particularly, users who received downvotes during community uncertainty responded with confusion. After a discussion of gender-neutral terms for "dude" and "bro," one user responded to their downvotes by saying, *"Edit: I'm confused about the downvotes because our disagreement is so small. If I were you, I would do a poll of all the people I am talking with to figure out what they all want to be called."* Here, downvotes initiate a larger debate about what gender-neutral terms the community prefers. This supports our finding in Section 5.2.1 as downvotes instigate users trying to convince one another of their stance, thereby creating new community norms.

Finally, users respond to downvotes by maintaining their stance. Contempt appears in threads where individual stance meets community uncertainty. In a thread about the definition on non-binary, a user was downvoted for pointing out that a common shorthand, NB, should only be used about 'non-black.' After a long debate that concluded with users stating, *"words can be used for different things in different contexts,"* the user who brought up the issue stated, *"Edit: It is disappointing to see so many queer appropriators."* This confusion and contempt may arise because of the strong connection between downvotes and rejection.

5.3.2 Education and awareness. Education and awareness occurs when a community informs a user about a specific community understanding, peer-reviewed research, or other gender-related resources that help the user understand what it means to be non-binary. We find education and awareness in community rejection of the individual and individuals changing their stance. These findings help to answer RQ3 by exemplifying how community discussions of education and awareness are used to enforce community norms and influence users to edit their posts.

Education and awareness is a community reaction to users who infringe on community norms. When one user complained about feeling overwhelmed by the amount of cisgender people who sometimes occupy transgender spaces, the community was quick to make the user aware that not all non-binary people consider themselves to be transgender. This increase in awareness led the original user to post, *"Edit: I've changed the wording of my post. My apologies for thinking that all nb people are trans."* User responses to education and awareness show users learning community norms and editing their posts to comply with them. Therefore, discussing education and awareness is a tool for communities to initiate users into the norms of the communities.

We also see users change their stance in response to receiving education and awareness. When one user posted about certain hormone replacement drugs being toxic over time, other users responded with web links to medical research and blogs disproving the original user's post. The original poster edited their post to say, *"I've been informed that what I said doesn't apply to all hormone replacement drugs."* As a result, community education and awareness is a mechanism by which community uncertainty turns into settled community norms.

5.3.3 Individual support. We find support to be an individual's strategy to give comfort to the user who posted the initial post. An overwhelming amount of individual support leads to community acceptance or individuals maintaining their stance in the face of community uncertainty. These findings help answer RQ3 by exhibiting how users edit their posts in response to community norms in discussions of support.

In the case of community acceptance, individual support manifests in the form of admiration. When one user shared an image of themselves in swimwear, another user responded, *"This type of swimsuit made me more euphoric than anything else before! Well done on finding something that makes you comfortable!"* The initial user, like many users that receive individual support, responded with gratitude, *"Edit: These comments are making me cry. Thank you for lifting my self-esteem, I've never been the most confident. Thank you!"* Individual support is, therefore, a means by which community acceptance occurs and community norms are reinforced.

Finally, individuals provide emotional support to users who face community uncertainty. When one user asked for advice after their doctor declined to perform top surgery on them, another user dedicated their editing behavior to guiding the other user, *"Edit: I am sorry. Some surgeons out there will absolutely do top surgery on people that don't take testosterone. Don't give up hope!"* Therefore, individual support is a care-based mechanism where the community gently assimilates a user into the community norms.

6 Discussion

Our findings demonstrate that, in addition to influencing individual engagement, online communities use common social media and platform-unique design features to create and maintain community-specific norms around acceptable behavior and language. We begin by addressing the similarities and inconsistencies between the quantitative results and qualitative findings. Then, we discuss the impact of community norms on individual identity development, community safety, and social media governance. Online communities of non-binary individuals offer a unique case study for these psychological and organizational topics because of the communities' narrow focus on identity, exposure to online harassment and offline harm, and loose hierarchy among community members.

Across our mixed-methods study we find two inconsistencies in our findings related to a disconnect between self-disclosure and personal experiences and the context-dependent effect of scores on posts and comments. Here, we discuss these inconsistencies, unpacking them to help us understand why they may be occurring.

Our quantitative results show how users are more likely to edit their post/comment with each additional self-disclosure made within a post/comment (see: Sections 4.1 and 4.2). However, our qualitative findings show how the communities we examined favored users who shared their personal experiences, with these experienced being developed through self-disclosures (see: Sections 5.1.3 and 5.1.1). One reason for this discrepancy could be due to the edit post feature being used to show gratitude to the community where the community signals acceptance to them (see: Section 5.1.4); a positive way for users to respond to community members through post edits.

Sharing personal stories in non-binary communities is therefore an implicit community norm that encourages community inclusivity and solidarity.

Our results also highlight how scores lead to different outcomes depending on the context. In our quantitative results, we demonstrate that an increase in score having a *negative* influence on the likelihood of a user editing their post (see: Section 4.1). However, we also find that an increase in score has a *positive* influence on the likelihood of a user editing their comment (see: Section 4.2). Therefore, low-scoring posts and high-scoring comments are more likely to be edited. These result are supported by our qualitative findings which show how downvotes are used to both create and reinforce community norms (see: Section 5.3.1). As such, norms may also exist for the types of community responses (such as scores) that are expected in different contexts in addition to the type of posts that are deemed acceptable.

Overall, we found edit reasons being used be users not just to explain edits, but to also reply to comments without requiring the user to draft a new message. Our qualitative findings support our quantitative finding that posts attracting more comments are more likely to be edited than posts with fewer comments (see: Section 4). The number of reply comments could be reflective of the community showing support and signaling acceptance to the original poster. In these cases, we find users editing their posts to communicate gratitude to the community and their acceptance. However, we also see the effect of comments on editing behaviors during community uncertainty, where comments are used to debate controversial topics. Here, users edit their post to directly respond to community criticism and to argue for their point of view, such as one user who received backlash after claiming that a user who posted, "why does this subreddit praise androgyny and aesthetics even though that is not what being non-binary is about," was incorrect in their assumptions about the community. After receiving backlash for a comment that the community considered to be, "further marginalizing the problem," the critical user edited their comment to say that they, "want to be less marginalizing and more engaged but don't know how," and shutting down the discussion. Therefore, comments play an influencing role in user editing behavior which in turn leads to the creation and enforcement of implicit community norms.

6.1 Governing Identity-Based Communities through Implicit Norms

Our results show how implicit community norms contribute to the governance of identity-based communities. Prior work highlights how implicit norms protect community members [37] and how norms differ among communities [14, 36, 39, 85, 106]. Our study expands upon this prior work by showing how community norms exist within the broader ecosystem of community and platform rules on Reddit. Rules and norms on Reddit have been explored concerning direct community moderation [38, 54], such as Chandrasekharan et al.'s classification of community norms through comments removed by moderators [16]. However, our results show that platform affordances, such as upvotes/downvotes, influence norm creation and maintenance and that these norms can be seen through user editing behavior. Moreover, out findings support and extend the work of Dym and Fiesler [37]. In their work, similar to ours, they show implicit norms may ignite conflict between community members. Our work presents the various nuanced ways norms are enforced through community consensus and contested through community uncertainty.

Building on this prior work, we propose a model (Figure 3) for understanding how norms are created and maintained in identity-based communities. Firstly, our findings show how norms are maintained where there is community consensus around the content of an individual posts in relation to embedded behavior and language within the post (1). Where community consensus exists, we see the community either forming consensus to accept, reject, or be impartial towards the user of the post, and/or the post itself (2). Mechanisms used to signal acceptance include upvotes and supportive comments, while rejection occurs through downvotes, and impartiality occurs as

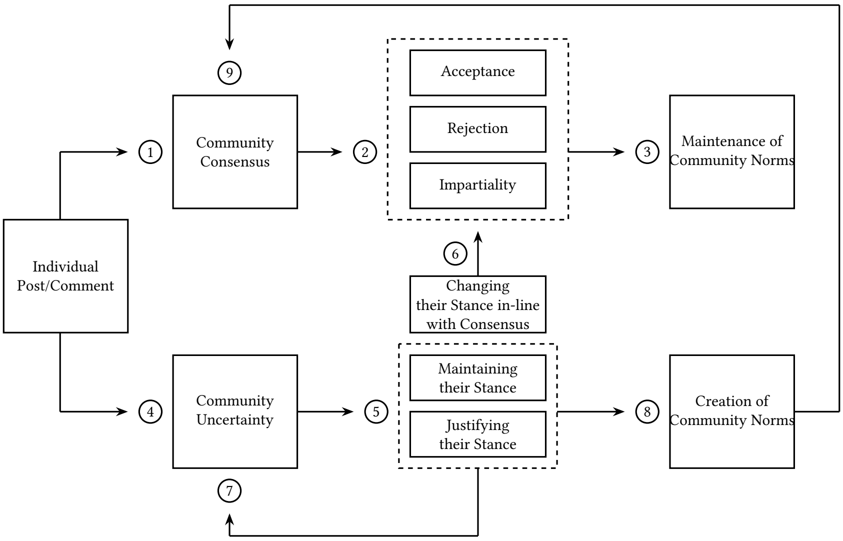

Fig. 3. Model of Implicit Norm as Governance in Identity-Based Communities

a result of little to no community feedback. The community may provide education and support when both rejecting a user and when showing impartiality towards a user. This suggesting that, in some cases, the community will educate a user to community norms which is aligned with prior work showing how members act to educate new users into a community who may be less aware of these implicit rules [6, 60]. These mechanisms are used to help enforce community norms, with the very act of enforcing community norms, regardless of norm's content, being used to promote safety within a community (3). This finding is reflective of the work of Dym and Fiesler who found community norms being used to help protect users within fandom communities [32].

Secondly, our analysis suggests that norms are created iteratively as many users refine their individual posts within and across discussions. This part of the model particularly responds to Devito et al.'s call to investigate how norms are formed online [27]. We find that these discussions occur as users maintain or justify their stance (5) in the face of opposition from other users due to the community uncertainty around the behavior or language embedded within a post. This finding also extends Fiesler and Bruckman's work which found that norms form through "observation, migration, and formalization" [37]. Our data does not support migration or formalization, due to the limitations of only studying a single platform, but we do find support for observation in the process of norm formation. Furthermore, we notice that the passive act of observation often leads to conversations that develop new community norms (8) that can result in the community developing consensus around that behavior or language (9). As users maintain or justify their stance, they engage in debates where users try to convince one another of the validity of their posts (7) or users change their stance to fall in line with a consensus view within the community (6). While it is difficult to recognize when these discussions precisely create a new norm, future work could directly involve community members in determining how implicit norms harden into community consensus

6.2 Developing Non-binary Identity through Community Norms

Online communities are a tool that enables non-binary users to "do non-binary." Conlin et al. found that non-binary people defined their gender experiences in terms of "(a) identity development,

(b) heterogeneous identities, (c) identity-expression divide, (d) invisibility and stressors, and (e) resilience and support" [19]. Our results show how the existence of non-binary communities on Reddit creates spaces for non-binary people to unpack these experiences and create a shared knowledge of how to thrive as non-binary, yet there are challenges in the governance of these communities.

We found implicit community norms supporting people in talking about identity-expression divides, invisibility and stressors, and resilience and support. Community members embraced people, through community acceptance, who separated their gender identity from their gender expression. In this way, community norms encourage users to continue discussions around the identity-expression divide and to do/redo/undo gender offline [23]. Discussions around invisibility and stressors also often surfaced through community rejection when newcomers to the community were unaware of the challenges non-binary people faced. Community norms can therefore act as a safety mechanism by which community members discern the intentions of a newcomer. If the member who used the phrase ‘socialized male’ does not understand their transgression, then the community knows that the member does not suffer the same invisibility and stressors and may not be welcome in the community. Discussions of resilience and support spanned across all themes. Resilience and support are built into the purpose of the communities we investigate, yet this result shows that community norms may act in parallel with explicit norms to encourage pro-social behavior.

The norms developed in the communities studied also raise concerns over identity development and homogenizing diverse non-binary identities. We uncover evidence suggesting that community members expect newcomers to have resolved their identity development, with a particular focus on aesthetics. As previously stated, this may relate to safety – it is difficult for community members to determine the authenticity of a newcomer. It is, therefore, easier for a community to welcome a newcomer that they deem “non-binary enough,” and we see strong community norms mediate those relationships. We also find that the communities struggle with the diverse ways that members express and experience their non-binary identity, particularly concerning intersectionality. The diversity of these identities may confound the process of norm-making even though we observe users try to settle community uncertainty. In this way, the heterogeneity of non-binary identities evades the creation of norms similar to the way that Barbee and Schrock find that non-binary people evade binary classification [9].

6.3 Implications for Community Governance and Design

Currently the edit reason feature has been adopted through the appropriation of the interface, yet we find that it is a useful mechanisms to help people respond to perceived violations of social norms, and to help the shaping of community norms through edits that maintain a users stance, or act to justify their stance. While many post edits were identified, the majority did not feature an edit reason, perhaps as this behavior is not well understood, especially by newcomers within the community. Recording a post/comment’s edit history could surface these reasons without requiring a user’s labor. Beyond technical challenges, the edit history could be used by adversarial users to resurface embarrassing edits. Ensuring that edit histories could not be taken out of context would therefore be an important design component of this potential feature.

Toward improving community governance, we show that editing behavior can be used as a way to understand how users’ create and enforce community norms. This builds upon prior work that measures community norms through user deletion and community moderation [16, 54]. Therefore, researchers, community members, moderators, and platform designers should look to editing behavior to get a more complete picture of governance within identity-based communities. Dym and Fiesler argue that norms could be made more explicit, yet there are challenges with this approach as making implicit norms explicit is likely to result in a shift towards more top-down approaches to governance around those behaviors [32], and could have unintended negative consequences for the community [17]. Yet, this approach does better support users who are new to a community who are less aware of the more implicit social norms that exist.

Prior work proposes machine learning approaches to moderation which learn from user deletions and community moderation [16, 54]. While these approaches could be applied to user editing behavior, we caution against completely deferring to automated tools that learn community norms from data within a community. However, automated tools could be used in conjunction with community engagement to create reports of predicted norms. These reports could then be shared in a community post and used to engage the community in discussions of how to make the community more accessible for genuine newcomers. This approach would include community members, and not just moderators [14], and create an archival record for newcomers to learn from.

Fiesler and Bruckman suggest that “it is a better strategy to use positive measures that reinforce being part of a community rather than negative measures that push people away” [37]. Rather than having implicit norms made explicit, integrating positive messaging around posts could help to reinforce being part of a community and guiding users towards prosocial behavior, whilst allowing for the freedom to explore identity and help reinforce and reshape community aspects of identity. One approach could be to embed proactive content moderation mechanisms that "nudge" users towards prosocial behaviors when drafting messages [102]. Allowing individual communities the ability to set their own “nudges" would be an important component of this design as prior work shows that norms differ across various online communities [14, 59, 85].

Determining the trustworthiness of a newcomer still remains a challenge for identity-based communities. Reddit currently does allow for communities to be public, private, or restricted. In the case of restricted subreddits, anyone can view the content of the community, but only approved users can make posts and comments. We encourage identity-based subreddits to activate this feature, so users can become accustomed to community norms before posting. However, this feature does increase moderator work as they become tasked with approving user access to make posts.

7 Conclusion

Users within online communities of non-binary users abuse the platform features of Reddit to influence and silence fellow community members. Where previous work explores how moderation explores how moderation affects marginalized users on Reddit, we tackle understanding how communities of non-binary users use the Reddit platform to govern identity-based communities through implicit norms. We first conduct a statistical analysis on over 2 million Reddit posts and comments to measure the effect of scores and self-disclosure on user editing behavior. We show that while every additional upvote *decreases* the likelihood of a post being edited, comments are *more* likely to be edited with each additional upvote–suggesting that posts with low scores are often edited in an attempt to gain more upvotes by fitting in with community norms. We then use qualitative methods to further investigate how community interactions and Reddit's platform features influence users' editing behaviors in communities of non-binary and gender non-conforming individuals.

We find that community norms exist around the types of posts and language that the communities deem acceptable for users to post/use. The boundaries of these norms are then the lines that the communities draw between acceptable and unacceptable language. We find that community norms are created, contested, and reinforced through the interactions between community-level and individual-level understandings of what it means to be non-binary. Understanding the mechanisms in which unspoken community norms are enforced, how users abuse platform features to coerce other users, and the effects of self-censorship on individual gender identity and expression, is the first step in encouraging developers to build online communities that are safer for and more inclusive of individuals existing outside of the gender binary.

Acknowledgments

We would like to thank the anonymous reviewers, Lisa Malki, and Antonis Papasavva for their insightful feedback. This project was funded by the UK EPSRC grant EP/S022503/1 that supports the Centre for Doctoral Training in Cybersecurity delivered by UCL's Departments of Computer Science, Security & Crime Science, and Science, Technology, Engineering & Public Policy.

References

[1] Almuhimedi, H., Wilson, S., Liu, B., Sadeh, N., and Acquisti, A. Tweets are forever: a large-scale quantitative analysis of deleted tweets. In *Proceedings of the 2013 conference on Computer supported cooperative work* (2013), pp. 897–908.

[2] Ammari, T., Schoenebeck, S., and Romero, D. Self-declared throwaway accounts on reddit: How platform affordances and shared norms enable parenting disclosure and support. *Proceedings of the ACM on Human-Computer Interaction 3*, CSCW (2019), 1–30.

[3] Andalibi, N., Haimson, O. L., De Choudhury, M., and Forte, A. Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (2016), pp. 3906–3918.

[4] Andalibi, N., Lacombe-Duncan, A., Roosevelt, L., Wojciechowski, K., and Giniel, C. Lgbtq persons' use of online spaces to navigate conception, pregnancy, and pregnancy loss: An intersectional approach. *ACM Transactions on Computer-Human Interaction (TOCHI) 29*, 1 (2022), 1–46.

[5] Andalibi, N., Ozturk, P., and Forte, A. Sensitive self-disclosures, responses, and social support on instagram: The case of# depression. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing* (2017), pp. 1485–1500.

[6] Arguello, J., Butler, B. S., Joyce, E., Kraut, R., Ling, K. S., Rosé, C., and Wang, X. Talk to me: foundations for successful individual-group interactions in online communities. In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (2006), pp. 959–968.

[7] Bailey, M., Dittrich, D., Kenneally, E., and Maughan, D. The menlo report. *IEEE Security & Privacy* (2012).

[8] Barak, A., and Gluck-Ofri, O. Degree and reciprocity of self-disclosure in online forums. *CyberPsychology & Behavior 10*, 3 (2007), 407–417.

[9] Barbee, H., and Schrock, D. Un/gendering social selves: How nonbinary people navigate and experience a binarily gendered world. In *Sociological Forum* (2019), vol. 34, Wiley Online Library, pp. 572–593.

[10] Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., and Blackburn, J. The Pushshift Reddit Dataset. *Proceedings of the International AAAI Conference on Web and Social Media* (2020).

[11] Braun, V., and Clarke, V. One size fits all? what counts as quality practice in (reflexive) thematic analysis? *Qualitative research in psychology 18*, 3 (2021), 328–352.

[12] Braun, V., Clarke, V., and Hayfield, N. 'a starting point for your journey, not a map': Nikki hayfield in conversation with virginia braun and victoria clarke about thematic analysis. *Qualitative research in psychology 19*, 2 (2022), 424–445.

[13] Burnett, G., and Bonnici, L. Beyond the faq: Explicit and implicit norms in usenet newsgroups. *Library & information science research 25*, 3 (2003), 333–351.

[14] Chancellor, S., Hu, A., and De Choudhury, M. Norms matter: Contrasting social support around behavior change in online weight loss communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), pp. 1–14.

[15] Chancellor, S., Pater, J. A., Clear, T., Gilbert, E., and De Choudhury, M. # thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing* (2016), pp. 1201–1213.

[16] Chandrasekharan, E., Samory, M., Jhaver, S., Charvat, H., Bruckman, A., Lampe, C., Eisenstein, J., and Gilbert, E. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proceedings of the ACM on Human-Computer Interaction* (2018).

[17] Cheng, J., Danescu-Niculescu-Mizil, C., and Leskovec, J. How community feedback shapes user behavior. In *Proceedings of the International AAAI Conference on Web and Social Media* (2014), vol. 8, pp. 41–50.

[18] Chong, T., Maudet, N., Harima, K., and Igarashi, T. Exploring a makeup support system for transgender passing based on automatic gender recognition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing*

Systems (2021), pp. 1–13.

[19] Conlin, S. E., Douglass, R. P., Larson-Konar, D. M., Gluck, M. S., Fiume, C., and Heesacker, M. Exploring nonbinary gender identities: A qualitative content analysis. *Journal of LGBT Issues in Counseling 13*, 2 (2019), 114–133.

[20] Craig, S. L., and McInroy, L. You can form a part of yourself online: The influence of new media on identity development and coming out for lgbtq youth. *Journal of gay & lesbian mental health 18*, 1 (2014), 95–109.

[21] Dame-Griff, A. *The two revolutions: A history of the transgender internet.* NYU Press, 2023.

[22] Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., and Potts, C. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web* (2013), pp. 307–318.

[23] Darwin, H. Doing gender beyond the binary: A virtual ethnography. *Symbolic Interaction 40*, 3 (2017), 317–334.

[24] Darwin, H. Challenging the cisgender/transgender binary: Nonbinary people and the transgender label. *Gender & Society 34*, 3 (2020), 357–380.

[25] Das, D., and Semaan, B. Collaborative identity decolonization as reclaiming narrative agency: Identity work of bengali communities on quora. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022), Association for Computing Machinery.

[26] Das, S., and Kramer, A. Self-censorship on facebook. In *Proceedings of the International AAAI Conference on Web and Social Media* (2013), vol. 7, pp. 120–127.

[27] DeVito, M. A., Walker, A. M., and Birnholtz, J. 'too gay for facebook' presenting lgbtq+ identity throughout the personal social media ecosystem. *Proceedings of the ACM on Human-Computer Interaction 2*, CSCW (2018), 1–23.

[28] Donath, J. S. *Identity and Deception in the Virtual Community.* Routledge, 1998.

[29] Dosono, B., and Semaan, B. Moderation Practices as Emotional Labor in Sustaining Online Communities: The Case of AAPI Identity Work on Reddit. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), ACM.

[30] Dosono, B., and Semaan, B. Decolonizing tactics as collective resilience: Identity work of aapi communities on reddit. *Proceedings of the ACM on Human-Computer interaction 4*, CSCW1 (2020), 1–20.

[31] Dym, B., Brubaker, J. R., Fiesler, C., and Semaan, B. "coming out okay": Community narratives for LGBTQ identity recovery work. *Proceedings of the ACM on Human-Computer Interaction 3*, CSCW (2019), 1–28.

[32] Dym, B., and Fiesler, C. Social norm vulnerability and its consequences for privacy and safety in an online community. *Proceedings of the ACM on Human-Computer Interaction 4*, CSCW2 (2020), 1–24.

[33] Ellison, N., Heino, R., and Gibbs, J. Managing impressions online: Self-presentation processes in the online dating environment. *Journal of computer-mediated communication 11*, 2 (2006), 415–441.

[34] Farnham, S. D., and Churchill, E. F. Faceted identity, faceted lives: social and technical issues with being yourself online. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work* (2011), pp. 359–368.

[35] Feuston, J. L., DeVito, M. A., Scheuerman, M. K., Weathington, K., Benitez, M., Perez, B. Z., Sondheim, L., and Brubaker, J. R. "Do you ladies relate?": Experiences of gender diverse people in online eating disorder communities. *Proceedings of the ACM on Human-Computer Interaction 6*, CSCW2 (2022), 1–32.

[36] Feuston, J. L., Taylor, A. S., and Piper, A. M. Conformity of eating disorders through content moderation. *Proceedings of the ACM on Human-Computer Interaction 4*, CSCW1 (2020), 1–28.

[37] Fiesler, C., and Bruckman, A. S. Creativity, copyright, and close-knit communities: A case study of social norm formation and enforcement. *Proceedings of the ACM on Human-Computer Interaction 3*, GROUP (2019), 1–24.

[38] Fiesler, C., Jiang, J., McCann, J., Frye, K., and Brubaker, J. Reddit rules! characterizing an ecosystem of governance. In *Proceedings of the International AAAI Conference on Web and Social Media* (2018), vol. 12.

[39] Gach, K. Z., Fiesler, C., and Brubaker, J. R. “control your emotions, potter” an analysis of grief policing on facebook in response to celebrity death. *Proceedings of the ACM on Human-Computer Interaction 1*, CSCW (2017), 1–18.

[40] Galupo, M. P., Pulice-Farrow, L., and Pehl, E. “there is nothing to do about it”: Nonbinary individuals' experience of gender dysphoria. *Transgender Health 6*, 2 (2021), 101–110.

[41] Gibson, A. Free Speech and Safe Spaces: How Moderation Policies Shape Online Discussion Spaces. *Social Media + Society* (2019).

[42] Gilbert, E. Widespread underprovision on reddit. In *Proceedings of the 2013 conference on Computer supported cooperative work* (2013), pp. 803–808.

[43] Gilbert, S. Towards intersectional moderation: An alternative model of moderation built on care and power. *Proceedings of the ACM on Human-Computer Interaction 7*, CSCW2 (2023), 1–32.

[44] Gilbert, S. A. " i run the world's largest historical outreach project and it's on a cesspool of a website." moderating a public scholarship site on reddit: A case study of r/askhistorians. *Proceedings of the ACM on Human-Computer Interaction 4*, CSCW1 (2020), 1–27.

[45] Goffman, E. The presentation of self in everyday life.

[46] Grimmelmann, J. The virtues of moderation. *Yale JL & Tech. 17* (2015), 42.

[47] Guyan, K. *Queer data: Using gender, sex and sexuality data for action.* Bloomsbury Publishing, 2022.

[48] Haimson, O. L., Brubaker, J. R., Dombrowski, L., and Hayes, G. R. Disclosure, stress, and support during gender transition on facebook. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing* (2015), pp. 1176–1190.

[49] Haimson, O. L., Dame-Griff, A., Capello, E., and Richter, Z. Tumblr was a trans technology: the meaning, importance, history, and future of trans technologies. *Feminist media studies 21*, 3 (2021), 345–361.

[50] Haimson, O. L., Delmonaco, D., Nie, P., and Wegner, A. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proc. ACM Hum.-Comput. Interact. 5*, CSCW2 (2021).

[51] Hui, J., King, J., Mcleod, C., and Gonzales, A. High risk, high reward: Social networking online in under-resourced communities. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023), pp. 1–12.

[52] Jaroszewski, S., Lottridge, D., Haimson, O. L., and Quehl, K. "Genderfluid" or "Attack Helicopter": Responsible HCI research practice with non-binary gender variation in online communities. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (2018), pp. 1–15.

[53] Jhaver, S., Boylston, C., Yang, D., and Bruckman, A. Evaluating the effectiveness of deplatforming as a moderation strategy on twitter. *Proceedings of the ACM on Human-Computer Interaction 5*, CSCW2 (2021), 1–30.

[54] Jhaver, S., Bruckman, A., and Gilbert, E. Does transparency in moderation really matter? user behavior after content removal explanations on reddit. *Proceedings of the ACM on Human-Computer Interaction 3*, CSCW (2019), 1–27.

[55] Jhaver, S., Frey, S., and Zhang, A. X. Decentralizing platform power: A design space of multi-level governance in online social platforms. *Social Media+ Society 9*, 4 (2023), 20563051231207857.

[56] Jones, R., Colusso, L., Reinecke, K., and Hsieh, G. r/science: Challenges and opportunities in online science communication. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (2019), pp. 1–14.

[57] Juneja, P., Rama Subramanian, D., and Mitra, T. Through the Looking Glass: Study of Transparency in Reddit's Moderation Practices. *Proceedings of the ACM on Human-Computer Interaction* (2020).

[58] Kairam, S., Brzozowski, M., Huffaker, D., and Chi, E. Talking in circles: selective sharing in google+. In *Proceedings of the SIGCHI conference on human factors in computing systems* (2012), pp. 1065–1074.

[59] Kasunic, A., and Kaufman, G. "At least the pizzas you make are hot": Norms, values, and abrasive humor on the subreddit r/RoastMe. In *Proceedings of the International AAAI Conference on Web and Social Media* (2018), vol. 12.

[60] Kiene, C., Monroy-Hernández, A., and Hill, B. M. Surviving an "eternal September": How an online community managed a surge of newcomers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), pp. 1152–1156.

[61] Klassen, S., Kingsley, S., McCall, K., Weinberg, J., and Fiesler, C. More than a modern day green book: Exploring the online community of black twitter. *Proceedings of the ACM on Human-Computer Interaction 5*, CSCW2 (2021), 1–29.

[62] Koshy, V., Bajpai, T., Chandrasekharan, E., Sundaram, H., and Karahalios, K. Measuring user-moderator alignment on r/changemyview. *Proceedings of the ACM on Human-Computer Interaction 7*, CSCW2 (2023), 1–36.

[63] Kraut, R. E., and Resnick, P. *Building Successful Online Communities: Evidence-based Social Design.* MIT Press, 2012.

[64] Lampe, C., and Johnston, E. Follow the (slash) dot: effects of feedback on new members in an online community. In *Proceedings of the 2005 ACM International Conference on Supporting Group Work* (2005), pp. 11–20.

[65] Lampe, C., and Resnick, P. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (2004), pp. 543–550.

[66] Matias, J. N. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences 116*, 20 (2019), 9785–9789.

[67] Mayworm, S., DeVito, M. A., Delmonaco, D., Thach, H., and Haimson, O. L. Content moderation folk theories and perceptions of platform spirit among marginalized social media users. *Trans. Soc. Comput. 7*, 1 (2024).

[68] McRoberts, S., Ma, H., Hall, A., and Yarosh, S. Share first, save later: Performance of self through snapchat stories. In *Proceedings of the 2017 CHI conference on human factors in computing systems* (2017), pp. 6902–6911.

[69] Musgrave, T., Cummings, A., and Schoenebeck, S. Experiences of harm, healing, and joy among black women and femmes on social media. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022), pp. 1–17.

[70] Naserianhanzaei, E., and Koschate-Reis, M. Do group memberships online protect addicts in recovery against relapse? testing the social identity model of recovery in the online world. *Proceedings of the ACM on Human-Computer Interaction 5*, CSCW1 (2021), 1–18.

[71] Nonbinary Wiki. Directory of online communities. https://nonbinary.wiki/wiki/Directory_of_online_communities.

[72] Nova, F. F., DeVito, M. A., Saha, P., Rashid, K. S., Roy Turzo, S., Afrin, S., and Guha, S. Understanding how marginalized hijra in bangladesh navigate complex social media ecosystem. In *Companion Publication of the 2020*

Conference on Computer Supported Cooperative Work and Social Computing (2020), pp. 353–358.

[73] Nova, F. F., DeVito, M. A., Saha, P., Rashid, K. S., Roy Turzo, S., Afrin, S., and Guha, S. " facebook promotes more harassment" social media ecosystem, skill and marginalized hijra identity in bangladesh. *Proceedings of the ACM on Human-Computer Interaction 5*, CSCW1 (2021), 1–35.

[74] Papakyriakopoulos, O., Engelmann, S., and Winecoff, A. Upvotes? downvotes? no votes? understanding the relationship between reaction mechanisms and political discourse on reddit. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023), pp. 1–28.

[75] Patel, D., Pendse, S., De Choudhury, M., Dsane, S., Kruzan, K. P., Kumar, N., Singh, A., and Warner, M. Information-seeking, finding identity: Exploring the role of online health information in illness experience. In *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing* (2022), pp. 263–266.

[76] Ren, Y., Harper, F. M., Drenner, S., Terveen, L., Kiesler, S., Riedl, J., and Kraut, R. E. Building member attachment in online communities: Applying theories of group identity and interpersonal bonds. *MIS quarterly* (2012), 841–864.

[77] Sarkar, C., Wohn, D., Lampe, C., and DeMaagd, K. A quantitative explanation of governance in an online peer-production community. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2012), pp. 2939–2942.

[78] Scheuerman, M. K., Jiang, A., Spiel, K., and Brubaker, J. R. Revisiting gendered web forms: An evaluation of gender inputs with (non-) binary people. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (2021), pp. 1–18.

[79] Schoenebeck, S., and Blackwell, L. Reimagining social media governance: Harm, accountability, and repair. *Yale JL & Tech. 23* (2020), 113.

[80] Schoenebeck, S., Ellison, N. B., Blackwell, L., Bayer, J. B., and Falk, E. B. Playful backstalking and serious impression management: How young adults reflect on their past identities on facebook. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing* (2016), pp. 1475–1487.

[81] Seering, J. 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 (2020), 1–28.

[82] Seering, J., Kraut, R., and Dabbish, L. Shaping pro and anti-social behavior on twitch through moderation and example-setting. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing* (2017), pp. 111–125.

[83] Seering, J., Ng, F., Yao, Z., and Kaufman, G. Applications of social identity theory to research and design in computer-supported cooperative work. *Proceedings of the ACM on human-computer interaction 2*, CSCW (2018), 1–34.

[84] Semaan, B., Britton, L. M., and Dosono, B. Military masculinity and the travails of transitioning: Disclosure in social media. In *Proceedings of the 2017 ACM Conference on computer supported cooperative work and social computing* (2017), pp. 387–403.

[85] Sepahpour-Fard, M., and Quayle, M. How do mothers and fathers talk about parenting to different audiences? stereotypes and audience effects: an analysis of r/daddit, r/mommit, and r/parenting using topic modelling. In *Proceedings of the ACM Web Conference 2022* (2022), pp. 2696–2706.

[86] Shuster, S. M. Generational gaps or othering the other? *Expanding the Rainbow* (2019).

[87] Sleeper, M., Balebako, R., Das, S., McConahy, A. L., Wiese, J., and Cranor, L. F. The post that wasn't: exploring self-censorship on facebook. In *Proceedings of the 2013 conference on Computer supported cooperative work* (2013), pp. 793–802.

[88] Sleeper, M., Cranshaw, J., Kelley, P. G., Ur, B., Acquisti, A., Cranor, L. F., and Sadeh, N. " i read my twitter the next morning and was astonished" a conversational perspective on twitter regrets. In *Proceedings of the SIGCHI conference on human factors in computing systems* (2013), pp. 3277–3286.

[89] Spiel, K. " why are they all obsessed with gender?"—(non) binary navigations through technological infrastructures. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference* (2021), pp. 478–494.

[90] Stone, A. L., Nimmons, E. A., Salcido Jr, R., and Schnarrs, P. W. Multiplicity, race, and resilience: Transgender and non-binary people building community. *Sociological Inquiry 90*, 2 (2020), 226–248.

[91] Tausczik, Y. R., Dabbish, L. A., and Kraut, R. E. Building loyalty to online communities through bond and identity-based attachment to sub-groups. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (2014), pp. 146–157.

[92] Taylor, J., and Bruckman, A. Mitigating epistemic injustice: The online construction of a bisexual culture. *ACM Transactions on Computer-Human Interaction* (2024).

[93] Taylor, J., Simpson, E., Tran, A.-T., Brubaker, J. R., Fox, S. E., and Zhu, H. Cruising queer hci on the dl: A literature review of lgbtq+ people in hci. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2024), pp. 1–21.

[94] Thach, H., Mayworm, S., Thomas, M., and Haimson, O. L. Trans-centered moderation: Trans technology creators

and centering transness in platform and community governance. Association for Computing Machinery.

[95] Treem, J. W., and Leonardi, P. M. Social media use in organizations: Exploring the affordances of visibility, editability, persistence, and association. *Annals of the International Communication Association 36*, 1 (2013), 143–189.

[96] Twyman, M., Keegan, B. C., and Shaw, A. Black lives matter in wikipedia: Collective memory and collaboration around online social movements. In *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing* (2017), pp. 1400–1412.

[97] Vijlbrief, A., Saharso, S., and Ghorashi, H. Transcending the gender binary: Gender non-binary young adults in amsterdam. *Journal of LGBT Youth 17*, 1 (2020), 89–106.

[98] Vitak, J., and Kim, J. " you can't block people offline" examining how facebook's affordances shape the disclosure process. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (2014), pp. 461–474.

[99] Walker, A. M., and DeVito, M. A. "'more gay'fits in better": Intracommunity power dynamics and harms in online lgbtq+ spaces. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), pp. 1–15.

[100] Wang, Y., Norcie, G., Komanduri, S., Acquisti, A., Leon, P. G., and Cranor, L. F. " i regretted the minute i pressed share" a qualitative study of regrets on facebook. In *Proceedings of the seventh symposium on usable privacy and security* (2011), pp. 1–16.

[101] Warner, M., Lascau, L., Cox, A. L., Brumby, D. P., and Blandford, A. "oops...": Mobile message deletion in conversation error and regret remediation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), pp. 1–13.

[102] Warner, M., Strohmayer, A., Higgs, M., Rafiq, H., Yang, L., and Coventry, L. Key to kindness: reducing toxicity in online discourse through proactive content moderation in a mobile keyboard. *arXiv preprint arXiv:2401.10627* (2024).

[103] Warner, M., and Wang, V. Self-censorship in social networking sites (snss)–privacy concerns, privacy awareness, perceived vulnerability and information management. *Journal of Information, Communication and Ethics in Society 17*, 4 (2019), 375–394.

[104] Weld, G., Zhang, A. X., and Althoff, T. Making online communities 'better': a taxonomy of community values on reddit. In *Proceedings of the International AAAI Conference on Web and Social Media* (2024), vol. 18, pp. 1611–1633.

[105] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. The fair guiding principles for scientific data management and stewardship. *Scientific data* (2016).

[106] Yang, D., Yao, Z., Seering, J., and Kraut, R. The channel matters: Self-disclosure, reciprocity and social support in online cancer support groups. In *Proceedings of the 2019 chi conference on human factors in computing systems* (2019), pp. 1–15.

[107] Yilmaz, G. S., Gasaway, F., Ur, B., and Mondal, M. Perceptions of retrospective edits, changes, and deletion on social media. In *Proceedings of the International AAAI Conference on Web and Social Media* (2021), vol. 15, pp. 841–852.

[108] Yoo, E. "i can't just post anything i want": Self-management of south korean sports stars on social media. *International Review for the Sociology of Sport 57*, 3 (2022), 477–494.

[109] Zhang, B. Z., Liu, T., Corvite, S., Andalibi, N., and Haimson, O. L. Separate online networks during life transitions: Support, identity, and challenges in social media and online communities. *Proceedings of the ACM on Human-Computer Interaction 6*, CSCW2 (2022), 1–30.

[110] Zhao, X., Salehi, N., Naranjit, S., Alwaalan, S., Voida, S., and Cosley, D. The many faces of facebook: Experiencing social media as performance, exhibition, and personal archive. In *Proceedings of the SIGCHI conference on human factors in computing systems* (2013), pp. 1–10.

A Appendix

Name	Description	Examples (paraphrased)
Added Correction in Response to Replies	When a user changes their opinion or text to answer other users. Messages with this code often contradict the initial message or include spelling and grammar fixes.	"I genuinely had no idea that the term ’gender critical’ meant something so bad."
Added Extension in Response to Replies	When a user includes more message-related details to answer other users. Messages with this code expand the context and content of the original message including commenting on content moderation or the message’s tone.	"I’ll add my response to someone else’s comment because I keep seeing references to this one in other comments."
Added Reflection in Response to Replies	When a user meditates over a message or thread to answer other users. Messages with this code frequently summarizes a message, include keywords such as "think" or "feel," provide support to other users, or include meta-commentary on a user’s own message."	"The discussion is ok during the first few comments, but it quickly gets worse."
Added Standalone Correction	When a user changes their opinion or text of their own volition. Messages with this code often contradict the initial message or include spelling and grammar fixes.	"My message sounds hostile. I understand that being called ’it’ is traumatic for some people, but in the end, your pronouns are yours, not theirs."
Added Standalone Extension	When a user includes more details of their own volition. Messages with this code expand the context and content of the original message including commenting on content moderation or the message’s tone.	"Wearing an ascot feels much more comfortable than wearing a necktie because the silk tie is positioned between your collar and neck."
Added Standalone Reflection	When a user meditates over a message or thread of their own volition. Messages with this code frequently summarizes a message, include keywords such as "think" or "feel," provide support to other users, or include meta-commentary on a user’s own message.	"I apologize; I realize this isn’t precisely the kind of answer you were hoping for, but I reasoned that any response was better than none."

Table 8. Codebook for Edit Reasons

Received July 2024; revised October 2024; accepted December 2024