Making Content Moderation Less Frustrating

How Do Users Experience Explanatory Human and AI Moderation Messages

Erik Calleberg

School of Natural Sciences, Technology, and Environmental Studies.
User Experience and Interactive Media Design, Master's Programme, 30hp, P4415
Södertörn University
Huddinge, Sweden

ABSTRACT

This is a study where social media users' reactions and interpretations to different types of content moderation messages. The focus was to try and determine what reactions explanations, explanations by an AI, or explanations by human moderators had on social media users. The goal of the study was to have this study become a pilot study for future research to find a solution to fair and transparent content moderation. The main research question sought to find out what ways do user attitudes differ from content moderation messages if the moderator is human or an AI. It was found that users react more strongly against AI moderation and decision making, showing higher rates of frustration. Providing a reason for moderation increases fairness and transparency rate, regardless of human or AI delivering the decision.

2021, June 10, Huddinge, Sweden.

CCS CONCEPTS

• Human-centered computing→*Human-computer interaction (HCI)*; *Social media*.

KEYWORDS

Content moderation, user experience, AI, algorithmic experience

ACM Reference format:

Erik Calleberg. 2021.Making Content Moderation Less Frustrating: How Do Users Experience Human and AI moderation.

1 INTRODUCTION

Social media is enormous, too big to limit the spread of misinformation and hate speech. Twitter grosses over 6 000 tweets per second, totaling 500 million tweets a day [1]. To combat this, content moderation and active removal of content are conducted by humans and automatic moderators with different terms of conditions depending on the social media platform [2, 3]. However, the sheer volume of posts and data makes it hard for Facebook to keep its platform free of harmful content. The scale of these platforms is also a factor, where the limited resources of moderators have to work hard against a tide of posts [4]. To stop the growth and spread of malicious content, the owners of social media platforms rely heavily on automatic and algorithmic to assist their human moderators [2]. Groups like anti-maskers and other extremist organizations have made significant negative impacts on society, like the storming of the Capitol that also got Donald Trump banned on Twitter because of his violence-encouraging tweets [5].

Explainable AI (XAI) has been used recently to explain how banks calculate credit scores and loan applications. XAI provides promising inspiration for how an XAI in content moderation explanation could be used [6]. Industrially, XAI has seen various uses, ranging from hiring, criminal justice, healthcare, and education [7]. There has been an increasing call for transparency and interpretability in these fields where XAI is not present as the impact of AI decision-making becomes more apparent. Removing user-made content is nothing to take lightly, as moderating users' posts equals censoring their voice on that platform. Sometimes users do not even get to know what rule they broke or why their post got removed. Implementing XAI in explanatory content moderation could increase transparency and help make social media a safer place.

The black box these platforms employ to keep their content moderation methods behind makes it difficult for end-users to find out exactly why their posts got moderated [8]. A black box means a part of a complex system's inner workings that is hidden. A recent study showed that proper explanation reduced the number of removed posts of a user in the future, while non-explanatory removals had the opposite effect [8]. That is why ethical considerations are highly recommended. User behavior changes depending on how and if they have explained why their post got removed. Providing users with an automatic explanation to increase the transparency of the content moderation process could also benefit the users' perception of the platform and help them understand the rules on the various social media platforms.

2 LITERATURE STUDY

This section provides an overview of the current discussions and state-of-the-art content moderation, and explainable AI topics. The goal is to present an advanced level of academic text regarding the evolution of the discussion and research of the said fields.

2.1 Content moderation

As mentioned in the introduction, social media is troublesome to keep free from harmful content due to the volume of users on the current platforms and how people can express racism and hate speech. Jhaver et al.'s research explores content moderation's impact, with transparency regarding content removal on Reddit by categorizing over 22 000 removal explanations given by subreddit moderators [8]. The study provided seven main observations of user behavior post-moderation [8]:

- The high past removal rate for the user is associated with lower odds of posting in the future.
- Greater explanation rates characterize reduced odds of posting in the future.
- Having a higher fraction of explanations offered through comments rather than through flairs is associated with an increased likelihood of users posting in the future.
- Explanations provided by human moderators, rather than automated tools, are associated with lower odds of moderated users posting in the future.
- The high past removal rate for a user is associated with higher odds of that user experiencing a post-removal in the future
- Greater explanation rates characterize reduced odds of post removals in the future.
- Having a higher fraction of explanations offered through comments, rather than through flairs, is associated with a decreased likelihood of users experiencing a post-removal in the future.

This shows that not only do explanations matter to users, in that some do not break the rule again, but that explanations provided by human moderations can contribute to lower odds of the users receiving moderation from posting in the future [8].

It is particularly interesting for this study that there is practical proof of human moderation having a better effect on users, which

is beneficial to this work. However, the study connects user behaviors to their activity through assumptions, which is a slight reliability issue. The authors connect their work to other research, like Kiesler et al. [9] who contribute to the field with insights that users learn social norms of a community, both in real life and online. They highlight the impact of removing user content, mainly when that results in an increase in harmful content or the loss of a user. The systems need to be legitimate. The authors claim that moderation actions that do not censor the speaker will be met with less resistance, suggesting that transparent explanations do have the desirable effects of letting users have a better experience and for the community moderators to lower the future workload [9]. An underlying theme here is that online communities run by people rather than companies might have to rely on transparent and appreciated content moderation practices. Even if all communities and social media rely heavily on its users, Facebook has 2.8 billion active monthly users. Smaller online communities need to be better to their members or leave [9]. Facebook is a megacorporation with a reliable base of users, and it needs more than a few scandals for a considerate loss of users to concern them.

As introduced in the section above, Vaccaro, Sandvig, and Karahalios experimented with the design of appeals. As cited in their article [2, 10]: "*appeal procedures differ greatly, and the differences may have considerable impact on an individual's perception of procedural fairness.*"

The authors' experiment serves as the main inspiration for this paper, specifically how they formed their questionnaire, focusing on the user's decisions to appeal a moderation decision and experimental design [2]. The experimental design uses mockups, sketches, etc., to simulate a situation, and in this case, they are interface mockups of users receiving moderation on their content. The experiment included 182 participants with an average of 46 divided into the given four appeal conditions of no appeal option, appeal written by a human, written by AI, and behavioral appeal [2]. The participants were given a scenario of their Facebook accounts having been suspended. Experimental conditions vary whether and how they can appeal the decision if given the option to do so. All participant's appeal applications were told that their appeals were denied and their Facebook accounts were wrongly suspended. The authors then judged the responses from what they called FACT, described as user perceptions of fairness, accountability, the trustworthiness of algorithmic decisions, and feelings of control by using the questionnaire with open-ended questions [2].

They found that adding appeals to the interface mockup does not significantly improve the perceived fairness of the user compared to a no-appeal mockup, the opposite of their hypothesis that was backed by a rich theoretical background. Furthermore, it turned out that no appeal condition performed better than any other type of dependent measure [2]. That particular finding is exciting since it contradicts their theoretical background, which heavily suggests the opposite. It also contradicts the other works, even though their focus was not the same. Although, the ANOVA-results difference between the human written appeal, although small, and the behavioral algorithm is to be noted, see figure 1 [2].

	No Appeal		Appeal							
			written human		*written algorithm*		*behavioral algorithm*			
Measures	M[a]	SD	M	SD	M	SD	M	SD	F[b]	p
Fair	3.97	0.87	3.94	0.78	3.81	0.80	3.77	0.84	F(3,163) = 0.837	0.45
Trustworthy	4.43	1.48	3.86	1.60	4.01	1.43	3.84	1.48	F(3,163) = 2.253	0.08
Accountable	4.76	1.18	4.51	1.32	4.29	1.35	4.23	1.35	F(3,163) = 2.305	0.07
Control	4.42	1.07	4.21	1.00	4.37	0.99	4.17	1.16	F(3,163) = 0.644	0.58

Fig 1. Ratings of the Likert-scale questionnaire, showing a slightly higher score for no appeal on average than any of the appeal variants.

The authors divided the responses into four sections: Contesting the Fundamental Goals, Contesting Automation, Contesting the Opacity of the System, and Contesting Inconsistency.

In the second part of the experiment, all participants were told that Facebook incorrectly identified them as spreading misinformation. Contesting the Fundamental Goals showed that while 61 of the participants understand and agree that Facebook should make an effort to moderate their content to limit misinformation, 48 others, however, say that Facebook should not

moderate content at all [2]. Although the participants were made aware that their content had incorrectly been identified as sharing misinformation, many accepted the verdict. A few even agreed with the algorithm, stating that the machine maybe knew something they did not and accepted the removal [2]. However, many users did argue that the error is obvious. They may have posted misinformation as they only post family pictures or rarely post anything at all [2]. Interestingly, some participants were motivated that their opinions and personal opinions are not misinformation. This suggests that opinions could not be classified as misinformation and that only claims of facts can [2]. These insights might be essential to remember in this work if the survey comes with similar findings. It exemplifies rigor and merit to both Vaccaro, Sandvig, and Karahlios [2], and this work is presented in this paper.

Vaccaro, Sandvig, and Karahalios [2] concluded in their report that none of their designed appeal choices improved the user's FACT perceptions compared to the "no appeal" variant, even if they were made to mimic Facebook's current design. The author suggests that their designs did not achieve the benefits that their research promised and that future work should seek improved designs. The users contest the decision given to them and about the issue itself of the goal of moderating content, automation in moderation, and inconsistency of the entire system. As users feel that their decisions do not affect the appeal system or give them a righteous chance, they mention leaving the platform [2]. This indicates that the actual problem is banning a user right out since the appeal versions themselves do not matter as much as the theoretical background made it out to be.

There are some dangers and harms with content moderation itself, which no doubt depends on the target of moderation as it depends on the moderator. There are numerous examples of marginalized communities getting censored on Facebook. One example is, after Israel arrested more than 200 Palestinians on charges of incitement on social media, Facebook removed 95% of the officials' removal requests after an agreement with Israeli officials. Facebook has denied the existence of these agreements. However, seven days after the alleged agreements with the Israeli officials, profiles of 7 Palestinians were removed [11]. Facebook has also been found banning LGBTQ+ ads after labeling them as political. LGBT groups are suspicious of Facebook's motives, with Facebook stating that they apologize for mistakenly banning ads, while some were deemed political. The organizations and companies making the ads suspect aversion from Facebook to show the ads to users [12].

The dangers of content moderation also concern the moderators themselves. Moderators on Facebook are underpaid and exposed to the horrors of the world wide web. Their job is to keep beheadings, bestiality, and child pornography off their platforms [13]. An anonymous moderator who worked for Facebook as an online moderator described his job as following [14]:

"There was literally nothing enjoyable about the job. You'd go into work at 9am every morning, turn on your computer and watch someone have their head cut off. Every day, every minute, that's what you see. Heads being cut off."

The moderator was paid 15$ an hour for removing terrorist content off Facebook. Media, profiles, and groups would either be flagged by users or algorithms to review, and the anonymous moderation team would decide if the content needed to be removed or escalated. Psychologists deem that the emotional toll of reviewing extreme imagery can be punishing. Even though workers should have been trained for extensive resilience in a two-week program before employment training, six months, and have been given access to counseling and the same support first responders are offered. The testimony given by workers might indicate that it is not enough as some workers have nightmares and seek to visit psychologists every day [14]. The work itself for content moderators is not straightforward either; hate speech has been reported to be challenging to deal with since they have to analyze the potentially lengthy content in its entirety to determine context and message [13]. This gets another layer of complexity when the work is outsourced in other parts of the world for the

workers who get barraged with racism, hate, and obscenities of other cultures [13].

2.2 Explainable AI (XAI)

AI has in recent years been called for to alleviate the moderators, to help cull the sheer volume of posts that get reported every day by tech industry leaders and policymakers. They hope that it will protect us all against harmful content such as hate speech and harassment to terrorist propaganda in a few years. A significant issue called the "restraint problem, " described as how filtering in social media will threaten freedom of speech before any reporting of users [15]. After upload, applying filters to all user-content instantly breaks the human rights framework articulated in 1948 by the UN and treaties formed in 1996. Any limitations of freedom of expression need to be provided by law, have a legitimate aim, and are required to have necessary and proportionate measures of achievement [16, 17]. Other, more logical problems with prior filtering are that it is impossible to restrict content before posting. Having filtering at upload reveals how the AI works, allowing the uploaders to understand the system. In keyword and hash matching filtering, users will game the system using abstract symbols, writing misleading titles, etc. What if a user posted something with the purpose of comedy, irony, or spreading awareness of an issue? Facebook had already shown problems with handling context in moderation when a mother outed a man calling her black son the n-word, which led to Facebook removing the post after 20 minutes for breaking company standards [18].

The call for ethical and efficient moderation might find a solution in explainable AI (XAI) as AI has been getting more and more integral in the role of everyday life, not to mention AI-based solutions in banks, marketing, security, and healthcare [7]. AI researchers and practitioners have put more attention on interpretability and explainability to make AI processes more transparent. One example where transparency is needed, as suggested by the content moderation section above, is the moderation process. Better transparency would help users learn more about how AI works, and researchers would better understand AI models at a scale. Because in current AI systems, there is often a Machine Learning (ML) centric model, and they are opaque, non-intuitive, and difficult to understand. Practitioners might be asking questions like "Why did you do that? Why not something else? When do you succeed or fail? How do I correct an error? When do I trust you?" [7]. The black-box AI does not only make it impossible for users to decipher the AI process, for better or worse. The goal of XAI is to make data predictable, clear, and transparent by, instead of having the data go into a black-box to AI product to decision/recommendation, the ML process will be revealed and made interpretable for the users and able for human inspection [7]. Other advantages with "Explainability," as Gade et al. describe it, are learning new insights, debugging, laws against discrimination, improving the ML model [7]. There are, of course, some significant challenges with successfully implementing XAI. There is a lack of a standard interface for ML models today that makes the implementation of explanations difficult. Explanations need variance depending on the user, field, and problem at hand, the algorithm needed for an explanation might depend on the case, type of ML model, and data format [7].

3 PROBLEM

The problem is that content moderation is getting increasingly automated and less transparent, which threatens to transpire into censoring user content if proper explanation to moderation is not provided. XAI appears to be a possible candidate to solve both of the predicaments; however, there has not been enough research for this implementation.

3.1 Research Question

In what ways do user attitudes differ from content moderation messages if the moderator is human or an AI? **a)** do attitudes differ (assessed with quantitative measure)?, and **b)** what do users think about this (assessed with qualitative means of assessment)?

4 METHOD

The study was selected to be quantitative with qualitative data elements through experimental design [19]. The work simulates users receiving a message of their content being removed from a website and elicits responses from the users when getting their content removed. In that sense, both seek to assess the users' opinions of frustration through reactions to prompts that can be quantitatively measured using Likert scales and assessed using qualitative means by using open text answers. Though, compared to qualitative methods where the usage of interviews and observation, Creswell & Creswell [19] would sort this study as quantitative. Data analysis is always daunting and time-consuming, regardless of the type of study. A qualitative study would demand several data sources and several steps to analyzing data due to the weight of the data, such as video, audio, images, and text. For example, an interview needs to be recorded, then transcribed, needs a thorough read-through, finds themes, groups the themes, and draws conclusions [19].

A quantitative survey will be used to compare user responses by using empathy probes/usability testing [20, 21]. A probe in the form of a digital mockup of a message that a user would get if they had the content they posted removed will be fabricated and used to get an understanding of their opinions on different variations of content-moderation messages see figure 2 [2]. The probe will be designed to be used in a quantitative survey to reach 50 responses as a minimum. The survey will utilize the Likert scale to measure user frustration to capture the intensity of user feelings toward the research in question. Otherwise, it might be challenging to guide the users to express their feelings if the questions are too open or lacking in structure [22]. The Likert scale, in this instance, will ask the respondents how frustrated the prompts make them, on a scale from 1, "not frustrating at all", to 9, "very frustrating" [23]. Without engaging questions and prompts, it will be troublesome to get any data at all as the survey and probe practically ask the respondents to imagine, to a certain degree, how it would feel to have their social media moderated.

The situational messages will vary from a select few social platforms and vary in their degree of explanation when motivating why the content was removed. The messages will also disclose if AI or a person made the explanation. The users will be asked to share their thoughts, and explain their reactions and opinions in free text forms. That particular data will no doubt be qualitative and analyzed as such (see Results & Thematic analysis). These types of testing are adequate for gauging real-time user reactions to the moderation messages, and having a mockup to test on gives the setting believable for the user to react as close as possible as if the situation would happen "in real life." Four different content moderation probes, M1-M4, will be fabricated similarly to figure 1 will be sent with the survey. Under each probe, a Likert scale will measure user attitudes and feeling intensity toward the probes to see if the transparency of explanatory messages and humanity matters in content moderation. Triggering responses will require prompts and follow-up questions to extract as many emotional responses as possible [21].

The different types of content moderation messages:

M1: A human moderated message, no explanation

M2: A human moderated message, explanation

M3: An AI moderated message, no explanation

M4: An AI moderated message, explanation

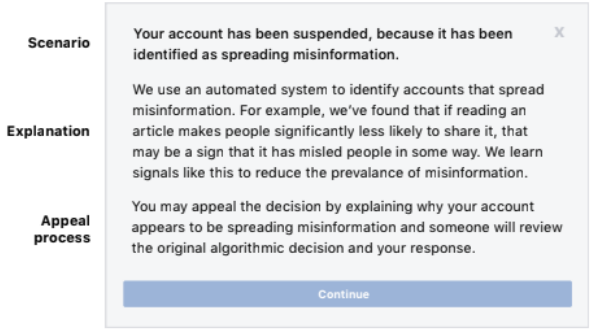

Fig 2. Example of a moderation message to test on users

Fig 3. A Likert scale

The qualitative data will be analyzed thematically, and they will be color-coded and apparent themes will be clustered together by doing a thematic analysis, explicitly using the method of affinity diagramming [21]. From this method, the users' perception of explainable AI and content moderation will be gained. These methods have been used previously to excellent results after doing interviews and surveys in another project. The methods are cheap, reliable if done systematically, and they are advantageous when trying to uncover themes, terminology, and frequency of occurrence [21]. A power analysis will be conducted to determine the quantitative answers' power to see if the sample size is sufficient. Then a Student's t-test will be conducted to determine if there are any significant differences between the four prompts [24].

The study's outcome will be insights into users' attitudes towards Ex AI, how they react to content moderation depending on a varying degree of explanation and whether a human moderation message is received differently from an AI-made moderation message.

A pilot study was conducted to test the questionnaires before publishing them to three students. They were sent a link to the questionnaire and were asked to test it, take time and come back with feedback. These answers were deleted before publishing the survey. The majority of the feedback was regarding the formulations of the questions, such as the prompts. Concerns were raised that without giving the participants an example of a moderated post, the user will have a hard time motivating how frustrated they will be. Another piece of feedback was the language in the prompts themselves that it was too professional and confusing. Specifically regarding the "appeal process." Lastly, some tips were given on making all questions mandatory.

The participants were recruited by social media, through posts on personal feed on Facebook, school groups on Facebook, Twitter, the subreddits r/SampleSize, and r/Assistance on Reddit, three different Discord channels.

4.1 Research Ethics

The institution mandates that the study adheres to GDPR as requested by Södertörn University. The only complications here were precisely the complexity of these documents. The data collected are user perceptions, age, gender, and location. All data will be removed upon receiving a passing grade. The following concerns are no dilemmas exactly, but possible annoyances that could impact the surveys. The survey consists of many free-text answers that could deter the participants from responding to it, as completion time could get higher, with reservations for people leaving it open and technical difficulties that occurred at one point where the Microsoft Forms servers went down temporarily. Another problem with the more qualitative questions is that the prompts and have the same questions to each of them might copy-paste their responses from the first one to the last. That would present their answers as low effort and possibly non-serious, which might call for not including them in the analysis. A major ethical consideration was not to do in-person user research due to the Covid-19 pandemic to help limit the spread of the virus.

5 RESULTS & THEMATIC ANALYSIS

This section aims to present the collected data and the results from the quantitative and qualitative data analysis.

The analysis followed card sorting and affinity diagramming methods to find patterns and themes among the longer text answers [21]. In total, there will be four sessions of affinity diagramming, one per prompt. Each of the prompt's qualitative responses will be thoroughly read through, apparent themes will be constructed, and the data will be color-coded to assign them a theme. To clarify, a response can be assigned to several themes if the situation demands it because otherwise, the importance of the qualitative data will be dismissed as longer texts where feelings are expressed can mean more than one theme [21].

5.1 Overview

The first part of the survey asked for their background, gender, age. The average completion time, according to Microsoft Forms, is 19 minutes and 1 second. Although it seems like a long time, some of the time could be attributed to a technical error that occurred at one point where the prompts did not load in. The long completion time could also mean that some of the participants struggled with the extended text answer-questions, spent a long time thinking about their answers, left the survey tab open, and forgot about it, skewing the average of the completion time.

Of the 50 respondents, 24 gendered themselves as women and 26 as men. An even divide between men and women should be beneficial for the study since all opinions about social media and content moderation are welcome and sought after in this context. The average age is rounded up to 29 years old, with a few at forty years old and older. As mentioned earlier, old or young people will not be treated differently. Nevertheless, an interesting angle for future research could be to investigate differences based on age.

The participants are mainly European. Many respondents have stated they are from Sweden. The others range from the U.K., North America, and the UAE. This is a classical western research project in that eastern users are not necessarily forgotten but hard to reach, and no effort has been made to reach further than Europe or the U.S. They have different social media platforms, and some have limited access to the internet [25]. The same goes for developing countries, they have not been purposefully ignored, but there have not been any attempts to reach them. The fourth question simply asked the participants if they use social media to ensure that their answers to the following questions could be taken seriously. There is no reason to explore user perceptions of content moderation if they do not use social media. If they answered 'no,' the questionnaire ended. 2 respondents answered 'No' on that question and therefore chose not to score the prompts.

Fig 4. Diagrams of question 5 and 6

Questions five and six seek to understand the users' habits and find how active they are and where. Thirty-two responded that they use social media daily and 15 hourly, which means that the chances of the users getting moderated in real life are higher the more they use the platform, but this does not show if they post content. Also, Facebook is the most used social media out of the respondents, this could help the study compare itself with Vaccaro, Sandvig and Karahalios's [2] study on Facebook's content moderation. Snapchat might be a controversial inclusion since it is mostly used for messaging. However, there are elements of a feed and public content through stories and under the Discover tab. The seventh question asks if the respondents have ever posted content on social media, and that could mean that they

have perceptions of the terms of service the various platforms impose on their users. This also indicates that they have seen, reacted to, gotten feedback on their posts, and possibly been moderated before. 96% of the respondents that answered 'Yes' on the fourth question, if they use any social media, answered 'Yes' for the seventh question which asks if they have ever posted something on social media. However, 64% of those who had posted content on social media do not consider themselves active posters on those platforms, which was asked in the following question.

At, and following the ninth question, starts the central part of the study of asking the participants to share their frustrations with content moderation and social media as a whole. Before arriving at the prompts, the participants are asked to recently share any frustrating experiences with social media and explain them. There is a mix of opinions by reading the responses, but at first glance, a prominent theme is annoyance with ads and 'toxicity'. Many (n=19) are tired of the fellow users on the platforms (*"Racists, homophobes, transphobes, and misogynists. Basically people who argue that others do not deserve human rights or respect, and even I have seen people advocating for violence and genocide. This frustruates me because hate speech should not be given a platform in mainstream social media"*.)[P21] and some (n=5) are really sick of the ads on social media (*"so many ads.. and recommended followers that I don't care about.*") [P15].

The four different prompts come with two statements that the participants are asked to rate the prompts from Strongly agree to Strongly disagree (Likert scale ref). The prompts, M1-M4, are presented below.

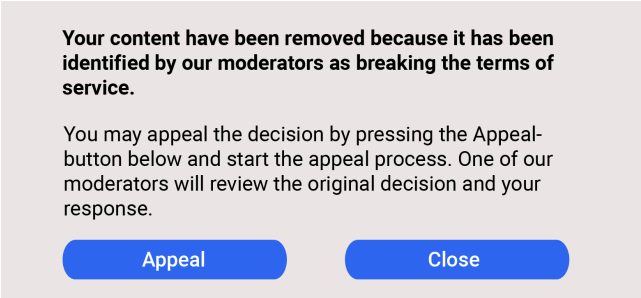

Fig 5. Prompt 1, M1.

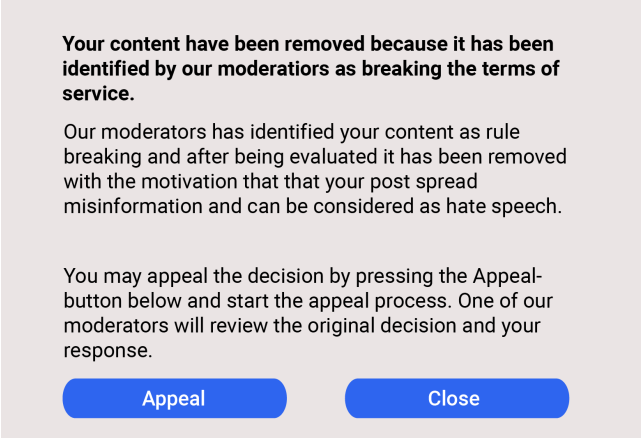

Fig 6. Prompt 2, M2.

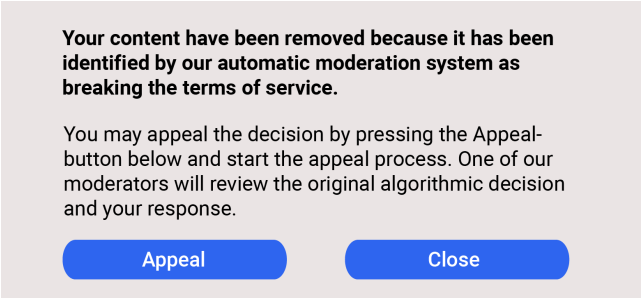

Fig 7. Prompt 3, M3.

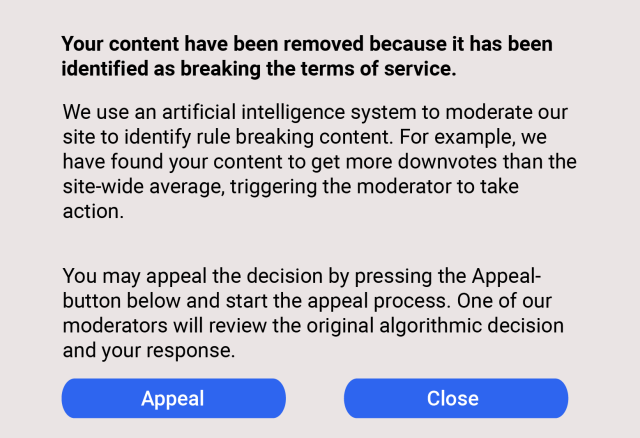

Fig 8. Prompt 4, M4.

5.2 Data from prompts

First, the results from the Likert-scale statements will be presented with box diagrams.

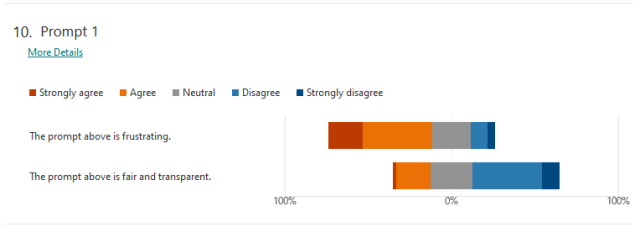

Fig 9. Results of M1.

The first prompt, a human moderated message with no explanation of the moderation, yielded the results seen in figure 4. 62.5% of the respondents agree or strongly agree that the first prompt is frustration, but only 52.1% disagree or strongly disagree that the prompt is fair and transparent.

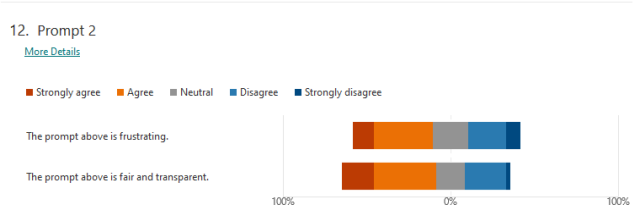

Fig 10. Results of M2.

The second prompt, a human moderated message with explanation, yielded the results seen in figure 5. Only 47.9% of the respondents agree or strongly agree that the prompt is frustrating, while 56.3% agree or strongly agree that the prompt is fair and transparent.

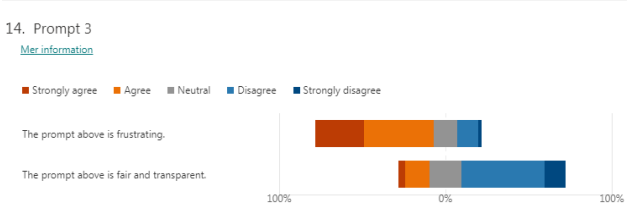

Fig 11. Results of M3.

The third prompt shows the respondents scoring 29.2% on Strongly agree and 41.7% on Agree as well as rating it very unfair and not transparent with 50% disagreeing with the statement and 12.5% strongly disagreeing.

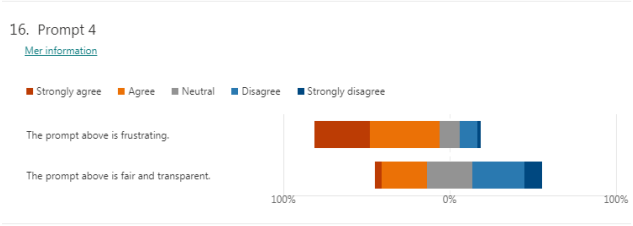

Fig 12. Results of M4.

The fourth, and last prompt, was rated similarly to the third and first prompt. 33.3% Strongly agreed that the prompt was frustrating, and another 41.7% agreed it was frustrating. The second statement yielded a more even spread of the answers with 31.3% reporting it as in any degree fair and transparent, and 41.7% as in any degree not fair and transparent.

In summary, these box diagrams illustrate that the results of M2, an explanatory moderation message with a human moderator, differ from the other prompts in how the users perceive it as more fair and transparent than any other prompt. However, it did not yield that notable variation concerning frustration. M1 and

M3-M4 appear similar in terms of the majority of users judging them to any degree frustrating and to any degree not fair or transparent. Even though M4 was aimed to be more explanatory than M3, the diagrams only indicate a small difference in user perceptions of the two statements.

The statements were quantified into numerical ratings. Strongly agree was given the value of 1, Agree 2, Neutral 3, Disagree 4, and Strongly disagree the value of 5. These values were then used to calculate the statistical power of the samples and to find if there are any significant differences in the prompts.

5.3 T.tests

T.test
Prompt1
0,00014
Prompt2
0,40419
Prompt3
0,00014
Prompt4
0,00011

Table 1. T.test results

A series of t.tests were conducted within the individual prompts to compare the responses of frustration versus fairness and transparency.

The t.test results align with the box diagram, showing that all prompts except M2, and proving that there is a significant difference between the responses of frustration and fair and transparent ratings. Presented in the table above are the p-values from the t.tests. It is apparent that M2 does not have a significant difference between the feelings of frustration and fairness and transparency as the p-value is higher than 0.05. Meaning, that the prompt was interpreted as both frustrating and fair and transparent. The other prompts did have a significant difference with p-values lower than 0.05, meaning that they agreed to the prompts being frustrating and disagreed with them being fair and transparent. They were all tested with an estimated power of 0.8 and an equal sample size.

5.4 Affinity Diagramming

Then the free text answers were thematically analyzed using affinity diagramming to capture emerging themes in the responses. The responses were read through, any emerging themes were written in a document, and then the answers were sorted into a fitting theme. If an answer did not fit into a theme made from the first read-through, it was assigned a new theme. The sorting was constructed so that if a long answer pointed to several themes, the text got scored with more than one because qualitative answers are not as clear as the Likert-scale scoring and can contain more than one opinion. Below is a table of the affinity diagramming session, where the texts have been categorized into cells that most closely correspond with a theme. The diagramming was done digitally in an Excel document and color-coded for clarity. The coding was systematically for each individual prompt. The free text question in the questionnaire was “Please describe your reasoning behind your answers above, why did you rate them as you did?”, for each prompt. The responses were submitted in English and Swedish, and any Swedish quotes in the response matrix will be translated to English in this report. Translated responses will be marked with a ‘ * ’.

Following is a description of the meaning of the themes found in the affinity diagramming session:

- Acceptance: The response indicates that they accept the verdict in the prompt.
- No reason given: The response indicates that the participants feel like they need a reason/explanation as to why they were moderated
- Fair and transparent: The respondent expresses that they deem the verdict fair and/or transparent (could also be ‘clear’).

- Design issues: Expresses problems with the design of the prompt itself
- Neutral: Expression of indifference and/or neutrality
- Frustration: The response clearly indicates or states frustration.
- Political: A response which debates whether something is right or wrong
- Problems with AI: The respondents expressed discontent with AI, algorithm, and automation
- Bad moderation reason: The respondents disagreed with the reason for the removal
- Unclear: The participant expressed uncertainty with the removal decision and/or process.

5.4.1 M1 - human moderation, no explanation

The answers to the first prompt can be summarized as frustrating uncertainty. n=14 wrote that they felt frustrated because they were not given a reason as to why they had their content moderated (*"It is frustrating because the reason behind the decision is not stated. Therefore, i would not say it is fair neither transparent.")*[P7]. n=23 expressed that they wanted a reason for why their post broke any rules and/or got removed, (*"The prompt lacks information regarding which content was removed and the reasoning behind the removal, therefore lacks transparency. It would've been less frustrating if it pointed out the specific term that was broken."*)[P38]. Another concern is that human moderators are biased when it comes to moderation, and that makes the decision faulty (*"They should look at the decision not at the information that was posted, manual checking of all posts sounds crazy because then the judging becomes subjective after the [moderators] opinions and taste"*)*[P23]. Interestingly, n=2 expressed discontent with AI even if the prompt explicitly state that a human moderator took action, one stating that the human moderator is not actually human and it is just a front for the platform to carry their agenda. ("*I understand that websites can have whatever terms they want. But it is often used with out actual humans seeing it and is just automated. And frankly it's often just part of a bullshit agenda")*[P38]. Finally, n=10 respondents accepted the verdict, several of them understanding that this needs to be done on the internet, (*"I've never posted something that breaks the term of service. If you wish to use a certain site you have to play by its rules."*)[P49].

5.4.2 M2 - human moderation, explanation

This prompt can be summarized as cautious acceptance. n=15 respondents accepted the explanation powered moderation, ("*I guess they write more reasons why the content has been removed*")[P13] and n=16 felt it was fair and transparent. These responses indicate that having an explanation with a moderation message will have a significant impact on the reaction of the users. Some expressed that the reason helped ease their frustration, even if it still was frustrating, a reason makes it feel fairer, (*"It is clear, but still frustrating to get"*)[P8]. n=12 uttered themselves politically, starting a discussion about both having humans as moderators and the reason for moderation stated in the prompt; hate speech. Some argued that hate speech is not a real thing (*"Hate speech isn't a real thing, and there's no clear definition. And one person's idea of hate speech could be completely different than another."*)[P28] and just a tool for silencing unwanted opinion, while others presented concerns about the neutrality of the moderator "*It feels obviously more legit to get a reason why your content has been removed. It's important to not spread misinformation and hate of crouse. Despite this, it's about whether I agree or not. Social media can often judge innocent posts as e.g "hate toward men" while apparent racist posts get to remain up. I don't trust that the moderator is neutral.*")*[P5]. The responses labeled as 'Unclear', felt that they needed a better explanation and that the prompt still did not explain why they broke the rules, (*"The prompt still doesn't explain how I broke any rules."*)[P40].

5.4.3 M3 - AI moderation, no explanation

A summary for this prompt shows a similarity to M1 with an added layer of discontent with bots making moderation decisions. An outstanding number is the n=26 participants expressing

discontent with bots as moderators. Many write that bots should not have the authority to make such decisions ("*No reason given again, and AI should not have such executive power.*")[P10], others simply seem to not trust them (*"Very annoying that an automatic screen could see what was ok and not ok to post something. I woud be very annoyed. I would have had a hard time to feel confident with a automatic moderation system looking through my content"*)[P15], saying that they are faulty and inaccurate ("*Algorhythms never get it right. They miss important things and they let important things slip through the cracks.*")[P21]. Ten participants felt that this was fair and transparent, several respondents stating that it was nice getting to know that there was an algorithm taking the decision (*"Its good to know that it first went through an algorithm and that its being viewed by a person afterwards."*)[P45].

5.4.4 M4 - AI moderation, explanation

A quick overview of the table shows that a majority of participants held spite the moderation reason n=28, and 13 having problems with the AI. The moderation reason read as follows: *"We use an artificial intelligence system to moderate our site to identify rule breaking content. For example, we have found your content to get more downvotes than the site average, triggering the moderator to take action."* Many participants explained the weak morality of moderating users because of what the other users think of their content, (*"Mentioning down votes sets us all up to a suppression of opinion, ideas and political discourse. We might not agree with a statement and down vote it, but that doesn't mean it should be silenced or forcefully removed. All opinions should be voiced for a healthy debate."*)[P36]. Despite this, n=12 deemed the prompt in some degree to be fair and transparent, a few (n=5) of them mentioning that giving them an insight into the process, makes them feel better, (*"As a user, I get to know a bit more about how the moderation system works, which puts me more at ease."*)[P24].

5.5 Summary

Affinity Diagramming	Prompt 1
Acceptance	10
No reason given	23
Fair and Transparent	5
Design Issues	1
Neutral	6
Frustration	14
Political	5
Problems with AI	2
Unclear	5

Affinity Diagramming	Prompt 2
Acceptance	15
No reason given	5
Fair and Transparent	16
Design Issues	3
Neutral	2
Frustration	12
Political	12
Dislikes moderation	1
Bad moderation reason	2
Unclear	5

Affinity Diagramming	Prompt 3
Acceptance	3
No reason given	13
Fair and Transparent	10
Design Issues	2
Neutral	4
Frustration	14
Political	1
Problems with AI	26
Unclear	4

Affinity Diagramming	Prompt 4
Acceptance	3
No reason given	3
Fair and Transparent	12
Design Issues	1
Frustration	6
Political	9
Problems with moderation	2
Bad moderation reason	28
Unclear	5
Problems with AI	13

Fig 12. M1-M4, results of Affinity Diagramming

All prompts invoked feelings of frustration in many participants, most likely because of the nature of receiving news stating the removal of their content. M2 was the prompt that users interpreted as the least frustrating and felt it the most fair and transparent of the prompts. It is most prominent in the box diagrams (see figure 6) where it is clearly visualized that M2 was rated as the least frustrating at 47.9%, and the most fair and transparent, with 56.3% agreeing to some degree. Prompt M1, M3 & M4 shared similar ratings of frustration and fairness, and transparency. M2 and M4 had the highest rates of participants deeming the prompts as fair and transparent, the answers being in a similarity of gratefulness to the explanations given. All prompts, except one with 4, had 5 answers coded as 'Unclear'. 'Unclear' means that the respondents reported that not enough information was given to them about the rules or how the system works and not that the prompt itself was unclear. Those answers were coded as 'Design issues'. An interesting find is, despite M1 clearly stating that it was moderated by a human, n=2 mentioned distrust with AI moderation (*"I understand that websites can have whatever terms they want. But it is often used with out actual humans seeing it and is just automated. And frankly it's often just part of a bullshit agenda"*)[P28]. This suggests that the participant distrust the message and AI in general to the degree of frustration toward moderation as a whole. They rated the prompt 'Strongly agree' when asked if they deemed it frustrating, and 'Disagreed' with it being fair and transparent. M3 and M4 are the two prompts that report moderation by an AI. M3, without moderation reason, triggered 26 participants to voice their opinions on algorithmic moderation. M4 also had a lot of complaints against it, but also on

the moderation reason. Overall, the results of the survey suggest the following:

1. Giving a reason for moderation increases fairness and transparency rate, regardless of human or AI delivering the decision.
2. Users want to see the post that has been taken down.
3. AI moderators taking moderation decisions result in a higher frustration rate in users.
4. Human moderator messages seem to be better received by the user.

6 DISCUSSION

As evident with the number of citations, and the inspiration for the design of the prompts, Vaccaro, Sandvig and Karahalio's [2] study were the main source for the study. Their design was altered into four different variations in this study to compare user interpretations and reactions between explanations and human vs AI moderation. Their design introduced the scenario of a user's content getting removed, and with four different versions of text that stated how the appeal would be handled [2]. In a sense, they are also looking at the interpretation and reactions of the users, however more on how their feelings from the 'impact' of the moderation will transgress when presented with the appeal process of either no appeal, written appeal with a human moderator, written appeal by algorithm or a behavioral appeal with an algorithm. Whereas in this study, the appeal process is not explained, and the designs instead reveal 4 different moderators, 2 with explanations and 2 without. As suggested in both studies, users hold a firm grudge against AI and algorithmic moderation. Of course, there can be some debate on the relation of the terms AI, algorithm, machines, automation, and bots and if they all really mean the same thing in science but the uses of the words will be based on the words of the users from the thematic analysis. Since the design of the prompts in this study specifically stated AI as a moderator, and for the purpose of this section, those terms will be treated as synonyms unless another context is clear. An interesting point of the debate are appeals. Vaccaro, Sandvig and Karahalios' [2] focus on that part specifically while this study puts the appeal process aside. In retrospect, it would probably be interesting to have asked the respondents which button they would push for each prompt. Whether that would be beneficial for the study in any way is debatable, since they were included in the design, and that some users stated that it was welcomed that the appeal button existed.

A major difference between the studies is the number of participants in the studies. Vaccaro, Sandvig and Karhalios [2] successfully had 182 participants, post-data clearing, for their first test, and the second test added 267 to bring the total number of participants up to 449. Comparing that with this study, 48 participants,post-data clearing, in one survey lowers the validity and reliability of the experiment. However, it should be said that Vaccaro, Sandvig and Karahalios [2] paid their participants for doing the survey, and that is a probable reason for how they got that many to complete their tasks. Even if the participant count is low in the survey, the qualitative free-text answers act as a counterweight to the low number of quantitative responses, as long as the qualitative responses fill some degree of a grounded theory, meaning that the responses start to repeat themselves when the respondents have said what there is to say about the matter (see Results). Examples of repetitive answers can be seen in the most frustrating prompts, such as M1 and M3 wherein M1 n=23 writes that they are frustrated that the reason for having their content removed is missing, and more prominently in M4, where 28 respondents criticized the reason for removal. It gets trickier when analyzing the second prompt because it scored as frustrating and yet fair and transparent. Here lies an opportunity for future research. The prompt suggests that content moderation reasons should be made and explained by human moderators. However, do the answers really fill a grounded theory? It was the only prompt that did not score above 20 answers on any of the themes in the affinity analysis, and the responses were very mixed. Though, the possible ground is the n=16 participants mentioning "better than the last one" or "here I get a reason" which exemplifies.

6.1 Likert scale and statistical analysis

There is some disagreement in education and research about whether you should run tests like the t-test or non-parametric hypothesis tests like the Mann-Whitney on Likert-scale data [26]. Research has suggested that both ways return similar levels of power, both type 1 error rates were never 3% over the nominal rate of 5%, even in the instance of unequal sample sizes. The differences found in said research were skewed, peaked, or multimodal distributions. However, since the data in this study are on a nominal ordinal interval, “Strongly agree” to “Strongly disagree” on a scale from 1 to 5 where 3 was neutral, it is problematic to find the average. Finding the average on a series of numbers is possible, but doing the same on “Agree”, “Strongly disagree” and “Neutral” is not. This makes it impossible to find the mean in the data [27].

Taking these arguments into consideration, it appears that the t.tests used on the data found in this study is not an optimal way of doing statistical analysis. However, the qualitative data, the box diagrams (see figures 9-12), supports the calculations of the t.tests conducted (see table 1), showing that M2 stands out in terms of fairness and transparency. Still, the t.tests do not stand on their own, which in reality, in other sorts of statistical analysis they can.

6.2 Design of prompts

As mentioned in the method section, the design was greatly inspired by Vaccaro, Sandvig and Karahalios’s prompt, the differences being the addition of a button and varying texts of moderation messages depending on the source of moderation. One thing to note is the reason for moderation in M4. It reads: *“We use an artificial intelligence system to moderate our site to identify rule breaking content. For example, we have found your content to get more downvotes than the site-wide average, triggering the moderator to take action.”* As the affinity diagramming showed, where 28 participants showed frustration toward the reason for moderation, the wording was unclear and the reason for moderation unfair. It was put in as a reason during the design of the post under the impression that this kind of moderation is a common practice. In reality, it is not. Reddit’s default preferences collapse/hide posts automatically if they have a score of -4 [28]. This makes it an improbable reason for moderation from the platform itself, particularly at Facebook, however, it is possible occurrences on online community moderated forums that moderators remove the content as they deem fit. In future research, the explanations made by AI need to be reworked and based on a used standard/model or supported suggestion but the models for making the AI decision process understandable for users are not there yet [7].

The design of M1-M3 successfully probed the respondents to react and was not flawed as the fourth prompt may have been. The responses were given by the participants aligned with what the prompt meant, meaning that they showed comprehension of the text and responded to the question. Overall, the design of the prompts was satisfactory and provided a chunk of interesting data for the study.

6.3 Analysis

The analysis was done through affinity diagramming. Usually, affinity diagramming is done together in a group to discuss the themes as the session is going on which helps find new themes, remove unnecessary ones, and come up with new ideas. In practice, the analysis was more a mix of content analysis and affinity diagramming as the data were coded and counted more than discussed and grouped [21]. Regardless, the qualitative analysis has provided valuable insights into how users interpret and react to content moderation.

6.3.1 Relating results to the theoretical background

See insights in the summary section under results.

1. Research has shown that explanations have an impact on user behavior and that not having explanations could risk users leaving the platform and/or continuing to post potentially harmful content [9, 8]. Content moderation also has an impact on user behavior in real life as online communities generally try to keep social norms intact. The survey supported these arguments as frustration ratings

were lower on explanatory prompts, M2 & M4. Explanations made by human moderators take time and effort, something that might not be possible by those working for social media platforms with the scale and moderation policies of Facebook [14, 13, 4].

2. This connects to the first insight, but it could be noteworthy to distinguish. The participants mentioned in their answers. across all prompts, that they would have liked to see their content that got removed. However, if done, makes it possible for posters with malicious intent to game the system by slightly modifying their content to let it through any automatic moderation. This connects to black-boxing, and in this case, it might be best not to show them the post, even if that would be transparent to do. As black-boxing has been a term used to describe the problem with opaque AI operating in various solutions spread through several fields and services, it is necessary for the platforms to keep their security intact [7,8]. What should not be black-boxed are the motivations for removal, but that could make this argument difficult for social media platforms to adopt. It comes down to a degree of explanation instead of an explanation. How detailed should the explanations be? What should the tone be? Casual or formal language? How much information about the moderation methods can be revealed? To answer that requires further research using different variants of explanations to find a golden standard for explanations that are deemed fair and transparent by users and at the same time easy to produce but, more importantly, still holding the integrity of the system intact.

3. It became clear that the respondents do not trust or like an AI as a moderator, not to mention taking such important decisions as removing user content [7]. This partly supports the way Facebook is handling moderation, having human moderations taking the decision of removing content after it has been flagged. In some subjects, it is hard to rule out human guidance, or even for moderation overall since context matters. As demonstrated earlier, Facebook's system has shown flaws in the past with how they conduct content moderation with several marginalized communities getting a sour aftertaste after what some claim are unmotivated censoring [11]. Another reason AI moderators might not work for the near future is determining what is hate speech or not, which already is up to the human moderators to decide [13]. Hate speech was something that was brought up in one of the 'Political' responses as something that is hard to define ("H*ate speech isn't a real thing, and there's no clear definition. And one person's idea of hate speech could be completely different than another*.") [P27]. Yet it is a reason for moderation on Facebook.

4. These insights relate to the previous one in that it supports, to some degree, that Facebook uses human moderators that take the decision to remove content [13,14]. But there are a number of liabilities to that process: bias, cost of mental health, and efficiency. Bias is a difficult one to handle. It could be argued that even an AI is biased with the opinions of the ones that create it but as exemplified with hate speech, some decisions are up to the opinion of the moderator. This becomes extra difficult when the moderation duties are outsourced to another part of the world, which takes away valuable insight and familiarity to jokes and context regarding potential hate speech. This argument was described by one of the participants in the free text response to M1: "*[...] *I also do not know who this "moderator" is, is it a person, is it an AI? If it's a person, her attitude towards the content will affect whether I can contradict this and if it's an AI, it's certainly a giant bias and I can not trust this either.*" [P5]. The same participant's response to M2 was more accepting ("**It obviously feels more legit to get a reason why one's content has been removed. It is important not to spread misinformation and hatred, of course. [...] I do not trust that the moderator is neutral**"). This is but one of the examples that explanations do matter to the many users in terms of fairness and transparency, the question is to what degree. The mental health of the moderation workers was not great and needed to be taken into account if explanations would have to be given to every deletion [14]. Additionally, custom explanations would not be possible, and to even make them remotely possible, they would have to be automated by utilizing XAI.

7 CONCLUSION

In this work, I have tested different types of content moderation prompts to see how users react and interpret them through a questionnaire, measuring with a Likert scale. To answer the research questions**,** I explore moderation messages similar to Facebook's design with 4 different variations for human moderation and AI moderation with explanation and no explanations. **a)** Yes, users react more strongly against AI moderation and decision making, showing higher rates of frustration. **b)** Users have expressed that giving a reason for moderation increases fairness and transparency rate, regardless of human or AI delivering the decision, users want to see their content that got deleted, AI moderators invokes higher levels of frustration in users, and human moderator messages seem to be better received by the user. In future research, the goal is to have a more substantial sample size for the Likert scale, the proper methods for statistical analysis, and a second focus on the design of the prompts to gain a better understanding of how the wording, colorization, and buttons make any difference for how user's reacted and interpreted content moderation from human and AI moderators.

8 REFERENCES

[1] InternetLiveStats 2021. Twitter Usage Statistics - Internet Live Stats. [online] Available at: <https://www.internetlivestats.com/twitter-statistics> [Accessed 15 Feb. 2021].

[2] Vaccaro, K., Sandvig, C. and Karahalios, K., 2020. 'At the End of the Day Facebook Does What ItWants': How Users Experience Contesting Algorithmic Content Moderation. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2), pp.1–22.

[3] Jiang, J. 'Aaron', Middler, S., Brubaker, J.R. and Fiesler, C., 2020. Characterizing Community Guidelines on Social Media Platforms. In: *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*. [online] CSCW '20: Computer Supported Cooperative Work and Social Computing. Virtual Event USA: ACM.pp.287–291. https://doi.org/10.1145/3406865.3418312.

[4] Gillespie, T., 2020. Content moderation, AI, and the question of scale. *Big Data & Society*, 7(2), p.205395172094323. https://doi.org/10.1177/2053951720943234.

[5] Herrman, J., 2021. What Was Donald Trump's Twitter? *The New York Times*. [online] 12 Jan. Available at: <https://www.nytimes.com/2021/01/12/style/trump-twitter-ban.html> [Accessed 15 Feb. 2021].

[6] Liao, Singh, Zhang, and Bellamy. 2020. <http://aix360.mybluemix.net/explanation_cust> [Accessed 15 Feb. 2021]. V Introduction to Explainable AI. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (*CHI EA '20*). Association for Computing Machinery, New York, NY, USA, 1–4. DOI:https://doi-org.till.biblextern.sh.se/10.1145/3334480.3375044

[7] Gade, K., Geyik, S.C., Kenthapadi, K., Mithal, V. and Taly, A., 2019. Explainable AI in Industry. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. [online] KDD '19: The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Anchorage AK USA: ACM.pp.3203–3204. https://doi.org/10.1145/3292500.3332281.

[8] Jhaver, S., Bruckman, A. and Gilbert, E., 2019. Does Transparency in Moderation Really Matter?: User Behavior After Content Removal Explanations on Reddit. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), pp.1–27. https://doi.org/10.1145/3359252.

[9] Kiesler, S.,Kraut, R., and Resnick, P. 2012. Regulating behavior in online communities. Building Successful Online Communities: Evidence-Based Social Design (2012).

[10] Leventhal, S. 1976. What Should Be Done with Equity Theory? New Approaches to the Study of Fairness in Social Relationships. (1976).

[11] Onlinecensorship 2021. *onlinecensorship.org — Offline-Online*. [online] Available at: <https://onlinecensorship.org/content/infographics> [Accessed 24 May 2021].

[12] The Advocate. 2018. *Facebook Marks LGBTQ Advertisements as Political, Blocking Them Again*. [online] Available at: <https://www.advocate.com/business/2018/10/07/facebook-marks-lgbtq-advertisements-political-blocking-them-again> [Accessed 24 May 2021].

[13] Roberts, S.T., 2016. Commercial Content Moderation: Digital Laborers' Dirty Work. *Dirty Work*, p.12.

[14] Solon, O. 2017. *Underpaid and overburdened: the life of a Facebook moderator*. [online] the Guardian. Available at: <http://www.theguardian.com/news/2017/may/25/facebook-moderator-underpaid-overburdened-extreme-content> [Accessed 24 May 2021].

[15] Llansó, E.J., 2020. No amount of "AI" in content moderation will solve filtering's prior-restraint problem. *Big Data & Society*, 7(1), p.205395172092068. https://doi.org/10.1177/2053951720920686.

[16] United Nations General Assembly., 1948 & 1996. *Universal Declaration of Human Rights*. [online] United Nations. Available at: <https://www.un.org/en/about-us/universal-declaration-of-human-rights> [Accessed 25 May 2021].

[17] United Nations Human Rights Committee 2011. *OHCHR | International Covenant on Civil and Political Rights*. [online] Available at: <https://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx> [Accessed 25 May 2021].

[18] Jan, T. and Dwoskin, E., 2017. A white man called her kids the n-word. Facebook stopped her from sharing it. *Washington Post*. [online] 31 Jul. Available at: <https://www.washingtonpost.com/business/economy/for-facebook-erasing-hate-speech-proves-a-daunting-challenge/2017/07/31/922d9bc6-6e3b-11e7-9c15-177740635e83_story.html> [Accessed 25 May 2021].

[19] Creswell, J W. & Creswell, D.,2018. Research design. Qualitative, quantitative & mixed methods approaches.5th Edition. Great Britain: Routledge.

[20] d.school 2018. *design thinking bootleg.* Hasso Plattner, Institute of Design at Stanford.

[21] Marsh, S., 2018. *User research: a practical guide to designing better products and services*. London: KoganPage.

[22] Wikipedia contributors, 'Likert scale', *Wikipedia, The Free Encyclopedia,* 21 February 2021, 18:28 UTC, <https://en.wikipedia.org/w/index.php?title=Likert_scale&oldid=1008126734> [accessed 8 March 2021]

[23] Ceaparu, I., Lazar, J., Bessiere, K., Robinson, J. and Shneiderman, B., 2004. Determining Causes and Severity of End-User Frustration. *International Journal of Human-Computer Interaction*, 17(3), pp.333–356. https://doi.org/10.1207/s15327590ijhc1703_3.

[24] Brownlee, J., 2018. A Gentle Introduction to Statistical Power and Power Analysis in Python. *Machine Learning Mastery*. Available at: <https://machinelearningmastery.com/statistical-power-and-power-analysis-in-python/> [Accessed 3 May 2021].

[25] Jiang, M. and Fu, K. 2018. Chinese Social Media and Big Data: Big Data, BigBrother, Big Profit?. *Policy & Internet, Vol 10, No 4. 2018.* doi: 10.1002/poi3.187.

[26] Stephanie, G. 2015. "Likert Scale Definition and Examples" From StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/likert-scale-definition-and-examples/

[27] Stephanie,G 2015. "Nominal Ordinal Interval Ratio & Cardinal: Examples" From StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/nominal-ordinal-interval-ratio/

[28] Reddit. 2021. *settings (reddit.com)*. [online] Available at: <https://www.reddit.com/prefs/> [Accessed 26 May 2021].