"At the End of the Day Facebook Does What It Wants": How Users Experience Contesting Algorithmic Content Moderation

KRISTEN VACCARO, University of Illinois Urbana-Champaign
CHRISTIAN SANDVIG, University of Michigan
KARRIE KARAHALIOS, University of Illinois Urbana-Champaign

Interest has grown in designing algorithmic decision making systems for contestability. In this work, we study how users experience contesting unfavorable social media content moderation decisions. A large-scale online experiment tests whether different forms of appeals can improve users' experiences of automated decision making. We study the impact on users' perceptions of the Fairness, Accountability, and Trustworthiness of algorithmic decisions, as well as their feelings of Control (FACT). Surprisingly, we find that none of the appeal designs improve FACT perceptions compared to a no appeal baseline. We qualitatively analyze how users write appeals, and find that they contest the decision itself, but also more fundamental issues like the goal of moderating content, the idea of automation, and the inconsistency of the system as a whole. We conclude with suggestions for – as well as a discussion of the challenges of – designing for contestability.

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

Additional Key Words and Phrases: content moderation; algorithmic experience

ACM Reference Format:
Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2020. "At the End of the Day Facebook Does What It Wants": How Users Experience Contesting Algorithmic Content Moderation. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 167 (October 2020), 22 pages. https://doi.org/10.1145/3415238

1 INTRODUCTION

As algorithmic decision making systems become both more prevalent and more visible, interest has grown in how to design them to be trustworthy, understandable and fair. Particularly as algorithmic decision making systems take on important governance functions, these systems need to ensure they establish legitimacy with users [53, 55, 99]. One research direction has focused on designing *contestability* into algorithmic decision making systems [42, 57, 103]. Recent work found that in high stakes domains (i.e., automated assessment tools for mental health), allowing practitioners to contest decisions can make systems more understandable, useful, and accountable [42].

However, designing for contestability is challenging [103]. Before designing for contestability, it is important to establish that the much simpler task of designing for contestation — for example, through appeals — will not address users' needs. While appeals may not provide the deep system engagement of contestability, where users "*understand, construct, shape and challenge model predictions*" [57], they nevertheless have practical and psychological benefits [67]. Appeals provide a

Authors' addresses: Kristen Vaccaro, kvaccaro@illinois.edu, University of Illinois Urbana-Champaign; Christian Sandvig, University of Michigan, sandvig@umich.edu; Karrie Karahalios, kkarahal@illinois.edu, University of Illinois Urbana-Champaign.

2573-0142/2020/10-ART167 $15.00
https://doi.org/10.1145/3415238

mechanism for users to correct individual decisions but also to provide feedback on the decision making process as a whole. As one example, Facebook eventually updated its algorithms, after facing considerable controversy and numerous appeals, when its *real name* policy algorithms systematically suspended accounts for Native American users with last names like "Lone Hill" or "Brown Eyes" [43].

Despite the potential of such appeal systems to help users co-construct decision making processes in sociotechnical systems, they are often poorly designed or non-existent. For many years, social media platforms did not allow appeals of content moderation decisions [106]. Facebook [105], Instagram [1], and Twitter [79] all recently introduced or updated their processes for appealing content moderation decisions. Still, recent work on social media content moderation systems found that users describe the appeal process as "speaking into a void" [76]. However, this prior work analyzed data from self-selected users with prior experience with content moderation systems.

In this work, we experimentally test the effect of multiple appeal designs on users' experiences of contesting these decisions, and understand how users experience contesting automated content moderation decisions. We test this via a large-scale online survey experiment with a representative sample, including those with no prior experience with content moderation. Participants engage with a relatable imaginary scenario where an algorithm suspends their social media account. The between-subjects experiment compares the effects of no appeal to three types of appeals: a written appeal evaluated by a person, a written appeal evaluated by an algorithm, and one "behavioral" appeal which provides an opportunity to alter the decision by changing one's behavior.

We study whether different appeal designs can improve users' perceptions of the fairness of the decision making process, because feeling one's voice has been heard is an important part of fairness judgements [68, 101]. A survey measures the perceived fairness, trustworthiness, and accountability of the decision, as well as participants' feelings of control. We abbreviate these measures – **F**airness, **A**ccountability, feelings of **C**ontrol and **T**rustworthiness – as FACT. We also analyze the appeals that users write qualitatively. How do users write appeals of algorithmically generated content moderation decisions: what do they choose to contest, and how do they argue?

We find that while users call for appeals, adding them does not improve FACT perceptions. Since our work finds that current appeal designs are perceived as less fair, trustworthy, and accountable than no appeal at all, clearly changes to design are required. In qualitatively analyzing how users contest an incorrect decision, we find that participants contest the decision itself, but also more fundamental issues like the goal of moderating content, the idea of automation, and the inconsistency of the system as a whole. We conclude with reflections on what our results suggest for designing for contestation and contestability in algorithmic decision making systems.

2 BACKGROUND

We briefly review the related work on appeals in general and appeals of social media content moderation, and connect this review to three experimental conditions in the experimental design.

2.1 Purpose and Impact of Appeals

Appeals address the concern that no decision making process will make ideal decisions every time. So when an incorrect decision is corrected, appeals naturally benefit that individual decision subject. And when a decision making process has systematic problems, appeals can form a mechanism for feedback from decision subjects. However, past work has shown that appeals can improve people's experience of decision making systems even in the absence of any control over the final decision or decision making process [101].

Outcome favorability strongly influences people's perceptions of a decision making system. People can have negative emotional reactions to an unfair situation even when it benefits them [98];

this is particularly true when they are harmed [60, 97]. But when an outcome is unfavorable, research has shown that adverse reactions can be mitigated by trust and feeling that procedures were fair and inclusive [9, 10, 70].

One particular aspect of procedural fairness, "procedural control" or "voice" addresses whether people feel that they have been able to express their arguments and opinion [68]. Prior research has shown that voice significantly effects both perceived fairness and satisfaction with decision making systems [68]. Even when people lack any real control over final decisions, feeling their arguments and opinion have been heard improves their feelings of fairness [101]. And the impact of this is particularly strong for marginalized or disempowered groups [11]. Thus, when appeals allow users to feel their voice has been heard, they can impact satisfaction and feelings of fairness, even when they do not improve the final decision.

For this reason, our experiment tests the effect of appeals specifically when the outcome remains unfavorable; that is, the appeal does not correct the error. If an appeal is able to improve perceptions, despite not improving the outcome, it has successfully engaged users' feeling of procedural control or voice. This also drives our choice of dependent measures: fairness, accountability, feelings of control and trustworthiness (FACT).

2.2 Design of Appeals

As Leventhal notes, "*appeal procedures differ greatly, and the differences may have considerable impact on an individual's perception of procedural fairness*" [67], where the appeal processes can range from highly informal (e.g., appealing an exam grade in a classroom) to highly formalized. We briefly overview the design of several common, structured appeals systems. Three of the most well-known include the legal, insurance, and credit scoring systems, which allow people to appeal some cases, claim decisions, and aspects of their credit history, respectively.

Appeals typically require the decision subject to explain why they believe a decision to be in error. For example, on Medicare appeals forms, the Department of Health and Human Services asks decision-subjects to explain: "*I do not agree with the determination decision on my claim because* _____" [78]. Some domains provide scaffolding to help people structure these explanations. For example, the Federal Trade Commision provides a sample appeal letter to send to credit reporting agencies [17]. The sample specifies what the letter should include: how to identify disputed items (e.g., that they should be identified by both source and type), what changes can be requested, what errors can be appealed, and so on.

As this last item suggests, many domans constrain what someone is allowed to appeal; legal appeals focus on arguments about trial procedure or interpretation of the law (not new evidence) [4] and credit scoring appeals only allow people to dispute inaccurate data points (e.g., the report includes a credit card from a different person), but not how the score was computed [18].

Most appeal processes are evaluated by a human decision maker. Indeed, in Europe, the GDPR has mandated that individuals subject to automated decisions must be able to "*obtain human intervention*" and contest those decisions [35]. And when dealing with algorithmic curation, many users report preferring a human in the loop [24]. However human review has challenges. If decisions are made frequently or are likely to be contested, large numbers of reviewers are necessary. Agents in credit reporting agencies, for example, review each error for only six minutes [45]. In addition, moderators must be trained to be consistent. To understand whether it is the appeal itself that provides benefits or the human oversight, our experiment tests appeals with human or algorithmic reviewers.

Past work has also argued that "*the perceived level of fairness will be reduced when there are barriers that deter dissatisfied individuals from lodging complaints. To be fair, the procedures safe and easy to use*" [67]. While our review of content moderation system appeal designs shows that

current systems are often fail this goal, our experiment aims to make the appeal process simple and straightforward.

2.3 Social Media Content Moderation

In recent years, social media content moderation has become the object of intense public [2, 23, 71, 94] and research interest [19, 33, 55, 73, 85]. Researchers have studied problems of content moderation from numerous directions: the inherent inconsistency and frequent inequity of decisions [33, 62, 72, 86, 91], the dangers the content moderation process poses to moderators [33, 84, 93, 95], the harms that content moderation decisions can have for social media users, particularly marginalized communities [13, 15, 48, 69], and even whether content moderation actually improves future behavior [49].

Our experimental scenario involves one type of content moderation: an account suspension. Suspending an account is the most stringent form of moderation and has serious impacts on users [76]: account suspensions can interfere with users' abilities to communicate with friends and loved ones, read the news, or advertise their business – and because many services use social logins, can prevent access to other platforms [76].

To prevent malicious users from gaming the system, social media platforms share little about the process for suspending accounts. Nevertheless, we briefly overview what is known. Accounts are typically given "strikes" for harmful behavior like posting child exploitative content, bullying, or spreading misinformation [27]. In some cases accounts may be suspended for a single strike, while others require multiple strikes. Facebook states "*we don't want people to game the system, so we do not share the specific number of strikes that leads to a temporary block or permanent suspension*" [27].

Instead, users must rely on the Facebook Terms of Service and Community Standards to understand the reasons an account could be suspended [25]. However, how those standards are operationalized by content moderators is hidden and changes frequently [76]. As a result, many users develop folk theories of how and why content or accounts are suspended, which often attribute other people's actions as the primary cause, rather than algorithmic decisions [76].

2.4 Appeals of Social Media Content Moderation

Since content moderation systems make incorrect decisions frequently and sometimes systematically [2, 13, 15, 43], platforms have faced considerable criticism. But for many years, social media platforms did not provide appeal processes for content moderation decisions or allowed appeals of only some kinds of decisions [106]. Facebook [105], Instagram [1], and Twitter [79] all recently introduced or updated their mechanisms for appealing content moderation decisions. Nevertheless, many users report not knowing how to appeal or run into problems during the process [76].

When users appeal an account suspension, Facebook asks them provide minimal information (email address, full name, and an image of an ID) to confirm their identity [25]. The "appeal" is the action of sending the request; the appeal system does not provide a space for a user's explanation [26]. While users report exchanges with Facebook moderators attempting to explain why the decision is in error (see for example [83]), users also report frustration with the lack of human interaction, describing the process as "*speaking into a void*" [76]. Despite this frustration, appeals have the potential to provide meaningful support for users to their voice has been heard.
Condition 1: *To test the optimal version of Facebook's current appeal design, we introduce an experimental condition where the user writes an explanation in their appeal, that will be evaluated by a person.*

However, given the almost two and a half billion Facebook users in 2019 [28], adjudicating decisions with human content moderators has its own challenges. As in the credit scoring example mentioned above, moderators lack time for careful decisions, particularly as social media platforms

have suffered from shortages of human moderators [93]. In addition, moderators must be trained to be consistent both with themselves and with others. This has led to complex manuals that attempt to structure the human decision makers for consistency, which can be hard for human reviewers to manage [58]. One potential solution is to automate the review process as well. But can procedural control or voice be felt when it is an algorithm, rather than a human, hearing one's argument?
Condition 2: *To test whether an algorithmic review can also achieve the beneficial effects of a human-reviewed appeal, we introduce an experimental condition where the user writes an explanation in their appeal, that will be evaluated by an algorithm.*

Finally, past work has argued that content moderation systems should place greater emphasis on user education over punishment [76]. While social media platforms have recently developed temporary bans [76], these protocols simply suspend a user for a shorter time without allowing a fundamental disagreement with the decision. We develop a new appeal design that prioritizes behavioral improvement over punishment. In it, the user can change their behavior during a short window before a final review. This appeal attempts to make the content moderation processes more educational, and see whether users feel more control than in written explanations of past behavior.
Condition 3: *To test a method that would prioritize behavior change over punishment, our experiment tests a new "behavioral" appeal.*

Finally, we seek to understand what people want to contest in algorithmic decision making systems. Past work has found that people wish to incorporate additional information to improve the fairness of algorithmic decisions [64]. This experiment addresses this, studying what users wish to contest in an appeal of an algorithmic content moderation decision.

In exploring these topics, we focus on two research questions:

RQ1 How does adding appeals change users' perceptions of FACT of an algorithmic decision making system?

a. Does adding an appeal increase FACT perceptions?

b. Do algorithmically-reviewed appeals improve FACT perceptions?

c. Do behavioral appeals improve FACT perceptions?

RQ2 What do users choose to contest in their appeals of algorithmic content moderation?

3 EXPERIMENTAL DESIGN

Our experiment uses a between-subjects design to study the impact of appeals on user perceptions of algorithmic content moderation systems. We briefly overview the experimental protocol, experimental intervention, and survey instrument.

3.1 Recruitment

Participants were recruited via Qualtrics to be representative of the United States in terms of age, gender, ethnicity, education level and household income. Having a Facebook account was also an inclusion criteria for participation. Full demographics are reported in the supplementary materials.

3.2 Experimental Protocol

A between-subjects design involves a scenario where participants' Facebook account has been suspended. Experimental conditions vary whether and how they can appeal the decision.

Our experiment exposes participants to a hypothetical scenario in which their Facebook account is suspended for spreading misinformation. Facebook was chosen because of its large user base;

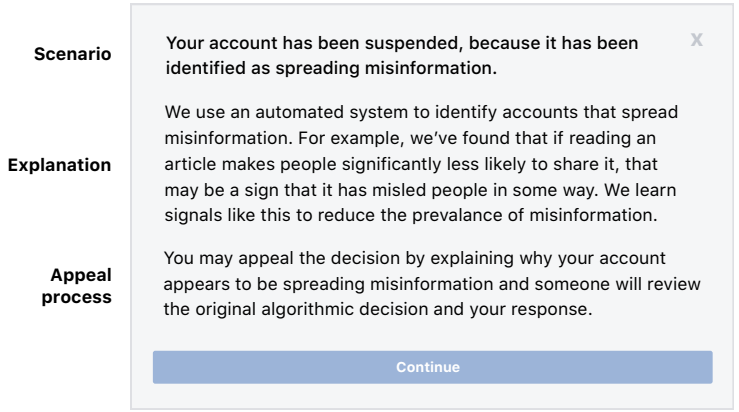

Fig. 1. Example of an interface mockup. This interface briefly introduced the scenario, explained the algorithmic decision making system and provided an appeal with *human* review.

with the exception of YouTube, "*no other major social media platform comes close to Facebook in terms of usage*" [36]. As noted above, having a Facebook account was as inclusion criteria for participation in the study. Only participants with an existing account could participate, to ensure that participants had a sense of how an account suspension would affect them.

As in Figure 1, an interface explains the decision and appeal process (if any). If an appeal is available, participants respond to the appeal prompt and whether they think their account will be re-opened. In all cases users are told that their appeal was denied and their account was *incorrectly* suspended. Most participants (78%) believe their appeal will succeed, but we want to address whether appeals can improve perceptions not only when a change improves the outcome but also when the decision is still unfavorable. We test this because in the face of unfavorable outcomes, perceptions of fairness and trust (which appeals might improve) are important moderating factors [9, 10].

Users then report their perceptions of the decision making system, addressing the FACT measures and a small set of open-ended questions.

The experiment uses a hypothetical scenario, which is very common in applied ethics and fairness research (e.g., [7, 44, 54, 80]). Third-party hypothetical scenarios are more common in human-computer interaction research [6], but first-person scenarios are so common in the fairness literature that some researchers studying such hypothetical scenarios noted, "*with a few notable exceptions, the study of perceived justice has been a 'first-person' undertaking*" [59]. They are also common in other domains, and work has found that there is no significant difference between hypothetical and field results in choice experiments [40].

3.3 Experimental Intervention

The experimental intervention is introduced via a mock-up of an interface. The interface has three sections (Figure 1). The top briefly introduces the scenario, that is, the cause for the account suspension. Next is a description of how the decision making system works, very similar to Facebook's own. Last is the available appeal process (if any) and the outcome of the appeal.

3.3.1 Scenario. Account suspensions are the most stringent form of moderation and research has primarily focused on their use as a form of punishment [30]. And in fact, account suspensions have serious impacts on users [76]. Thus it provides our experiment with a scenario with a violation of high *moral intensity* [50]. In particular, our scenario uses an incorrect suspension of a social media

Appeal	Description
None	We're always looking out for our users' safety and security, so you can't use Facebook right now.
Written, Human	You may appeal the decision by explaining why your account appears to be spreading misinformation and someone will review the original algorithmic decision and your response.
Written, Algorithm	You may appeal the decision by explaining why your account appears to be spreading misinformation and an algorithm will review the original algorithmic decision and your response.
Behavioral, Algorithm	Your account will be re-evaluated by an algorithm in one week based on your full activity history. If it is identified as spreading misinformation then, it will be suspended.

Table 1. Descriptions included in the interface. Each described a different type of appeal.

for spreading misinformation. Suspending accounts for spreading misinformation is both highly salient (e.g., [12, 16, 39]) and realistic (as found in pilot tests).

3.3.2 Explanation. A brief description explains how the decision making process works. Participants are shown a description drawn from Facebook's existing documentation on how they attempt to reduce the spread of misinformation [75]: "*We use an automated system to identify accounts that spread misinformation. For example, we've found that if reading an article makes people significantly less likely to share it, that may be a sign that it has misled people in some way. We learn signals like this to reduce the prevalence of misinformation.*"

3.3.3 Appeal Process. The final segment of the interface mockup describes the appeal process, if any. There are four experimental conditions in this experiment; the text descriptions used in the experiment for each condition are included in Table 1:

1. **no** appeal, as a baseline,
2. a **written** appeal, evaluated by a **person**,
3. a **written** appeal, evaluated by an **algorithm**, and
4. a **behavioral** appeal, evaluated by an **algorithm**.

The first condition – written appeal evaluated by a person – attempts to capture Facebook's existing process. While the current appeal process is not well publicized, users often report providing written explanations to moderators [76]. In our experiment, the user provides an explanation that they are told will be reviewed by a person (as shown in Figure 1), similar to the explanations users often report providing to moderators[1].

Two additional experimental conditions test automated review of the appeal, to address the potential challenges (of scaling and consistency) and whether human evaluation is necessary for users to feel procedural control and that their voice was heard. These conditions address whether *algorithmic* review can provide similar benefits, without requiring human reviewers. The second type of appeal mirrors the first: users write an explanation but are told an algorithm will review it.

[1] As noted in the Background section, Facebook typically also requires a photo identification, but we omit this to avoid collecting this personal information from participants.

The last conditions explores a possible improvement to appeal design, addressing calls for content moderation to be less focused on punishment [76]. The last condition is also evaluated by algorithm, but allows users to change their behavior, rather than provide a written explanation. Recent work has argued that content moderation systems should emphasize user education over punishment [76]. To help users learn and change how they act, social media platforms have introduced temporary bans [76]. But these protocols simply suspend a user for a shorter time, the decision still remains intact. Our experiment introduces a hybrid: the user can change their behavior during a short window before a final review. This version of an appeal provides users a warning and allows them to improve their on-going or even "clean up" past behavior before the final decision, encouraging behavior change. Using this appeal, we can test whether people feel more control when they can still engage with the system, rather than writing an explanation.

3.3.4 Outcome. In every case, after the intervention, participants are told that the appeal *failed* and that their account has been incorrectly suspended. We did this to ensure it was not just outcome favorability that improved perceptions, but actually engaging procedural control or voice. In addition, past work in social computing has shown that even malfunctioning mechanisms for control can improve feelings of satisfaction and control [102].

3.4 Survey Instrument

The full survey instrument is available as supplementary material; we summarize the design here. The survey begins with a consent form and concludes with a debriefing (recommended by the IRB) to reiterate to participants that they had no reason to anticipate an account suspension.

3.4.1 Prequestionnaire. Participants answered screener questions, including demographics and whether the user has a Facebook account. Participants without an account or who are unsure whether they have an account were excluded. Pre-intervention questions also addressed frequency of social media use and potential control variables (existing attitudes towards Facebook, propensity to trust social media platforms, and so on). A question about prior experience with social media account suspensions was also included, but placed after the experimental intervention, to avoid any potential priming effects.

3.4.2 Intervention. Users were asked to imagine that they see the experimental intervention's interface upon logging into Facebook. If an appeal was present, the user is asked to write a short text. Two conditions asked participants to write the explanation they would provide to Facebook. In the third they described their plan of what they would do in the following week. These writing tasks were designed to be similar to what real users report writing to Facebook after an account suspension (see for example [83]). Users in the no-appeal condition wrote nothing. After the appeal (if any) all users were told that their appeal was denied and their account was *incorrectly* suspended.

3.4.3 Main questionnaire. Users began with closed survey questions questions about four constructs: fairness, accountability, trustworthiness and feelings of control. Open-ended questions followed. These addressed their overall perception of the decision making process, Facebook's approach to fairness, and their trust in the reasoning provided.

Scales The questionnaire included shorter (4-5 question) scales for accountability, trustworthiness, and feelings of control, as well as a longer (17 item) scale for fairness. While all dependent measures were important, past work has shown that procedural control increases perceived fairness and, as noted by Flake et al., "measurement of broader or multidimensional constructs" – like fairness – "requires longer scales" [31]. Fairness is precisely such a construct; since justice is one of the central concepts in ethics, dozens of frameworks for fairness have been developed.

We developed the fairness scale to address a larger set of fairness frameworks than have been incorporated into prior work. Work in computing has focused primarily on distributive definitions [46], even work that examines how algorithmic definitions of fairness are perceived by the public [90]. However, when users evaluate algorithmic decisions they have different framings of fairness [65, 104]. In an experiment using a tool designed to create fair distributions of goods and tasks, Lee et al. discovered that users integrate multiple concepts of fairness and social context, going beyond the definitions of fairness that were designed into the decision making system [64].

To address this, we developed a fairness scale that addresses a broader set of fairness frameworks. It includes popular distributive notions [5] – that is, whether the final distribution of goods (or harms) is just – but also a variety of others, including more recent developments like restorative justice [8], which focuses on fully repairing the harms after a wrong.

While the full discussion of the development of this scale and its validation is outside of the scope of this work, it is described in the supplementary materials. In accordance with [31], we performed substantive review (including expert review), structural review (including factor analysis), and external review (measuring convergent and discriminant correlations with other scales meant to capture similar constructs).

For example, to address restorative justice, one item read: "*The Facebook decision making process gives a voice to the people who were most at risk from bad decisions.*" All 17 items are rated on a 7-point Likert scale from "strongly disagree" to "strongly agree". Two generic fairness scales from the literature were included, includng one well-validated scale used extensively in prior research [82]. This scale, the Reidenbach and Robin Multidimensional Ethics Scale (MES), was initially designed to capture different moral theories (e.g. utilitarianism vs. deontological theories), but does not capture different fairness frameworks.

The questionnaire also included shorter (4-5 question) scales for: accountability, drawn in part from prior work [38, 74]; trustworthiness, drawn from [81, 87]; and feelings of control, drawn from [81, 102]. Approximately 20% of the items were reverse coded.

Finally, a set of semantic scales measuring perceptions of fairness, trustworthiness, transparency and bias were repeated; the scales were shown first immediately after the intervention and again at the end of the experiment. These items provided a 7-point scale with endpoints of "fair" and "unfair," "trustworthy" and "untrustworthy," and so on. These were paired with an open-ended question about whether their opinion about the decision changed over the course of the survey; pilot participants shared that their perceptions of the decision improved, after realizing how many factors Facebook took into account.

Open-ended questions The survey also contained several sets of open ended questions. The first open ended questions were during the intervention, in which users wrote the appeal and why they would or would not expect that appeal to result in the re-opening of their account. The second asked the participant to explain their overall perception of the decision making process, after they had completed the FACT scales questions.

Participants also answered an open-ended question addressing whether they trusted the reason that was given for their account suspension, or whether they thought any other reasons might be responsible. Past work has found that users can feel targeted by suspensions [76]. And numerous researchers, reporters, and activitsts have noted cases where social media platforms make incorrect decisions systematically [2, 13, 15, 43]. For example, Facebook suspended many Native American users for violating their "real name" policy, because these users had last names like "Lone Hill" or "Brown Eyes" [43]. We designed this question to address the possibility that some users, particularly marginalized populations, might feel that they had been targeted, and to capture any folk theories of targeting for account suspensions.

3.4.4 Data sharing. The final question of the survey explained the benefits of data sharing and allowed participants to opt-in to share their data.

4 ANALYSIS

The quantitative results on the FACT scales are analyzed using a series of ANOVAs. Since the dependent measures are not strongly correlated ($r = 0.41 - 0.74$), we analyze each dependent measure separately. The quantitative results on the paired semantic scales to test for changes in perceptions over the course of the experiment are analyzed using a paired t-test.

The open-ended questions were analyzed via an iterative open coding. Initial codes were developed by the first author, discussed with the other authors, and iterated on to agreement. The analysis focused on the arguments participants developed about the decision making process. While we took an inductive approach, we did begin considering the focus on three broad areas (the input data, the method, and the final decision), based on the review of types of information allowed to be appealed in common appeal systems, as reviewed in the Background.

Since users continued reflecting on the decision making process and their arguments throughout survey, we analyze the data for all of the open ended questions as a single data set. For example, even in the no appeal condition participants often made arguments about the Facebook decision making process in their responses to how they felt about the process overall. For RQ2, we therefore also analyze data for all participants, even those in conditions with no or behavioral appeals.

5 RESULTS

Participants were recruited via Qualtrics to be representative of the United States in terms of age, gender, ethnicity, education level and household income. Participants were paid by Qualtrics for their participation, which took on average 20 minutes. Data cleaning removed responses that: 1) failed attention checks or 2) demonstrated nondifferentiation (straightlining) in their response. After data cleaning, responses from 182 participants were analyzed.

Results for RQ2 additionally include open-ended responses from a validation experiment of the questionnaire, described in the supplementary materials. These results include an additional 267 participants, similarly sampled by Qualtrics to achieve representation of the US population (full demographics of which are included in the supplementary materials). Thus the results for RQ2 include a total of 449 participants.

5.1 Impact of Appeals on FACT

The experiment includes results from 182 participants, with an average of 46 participants in each of the four appeal conditions: 50 in no appeal, 43 in written appeal, evaluated by a person, 43 in written appeal, evaluated by an algorithm, 46 in behavior appeal, evaluated by an algorithm. This experiment addresses our first research question:

RQ1 How does adding appeals change users' perceptions of FACT of an algorithmic decision making system?

Given the strong theoretical foundations of procedural control and voice, we had hypothesized that adding an appeal, particularly an appeal reviewed by a person, would improve feelings of FACT. And additionally, that a behavioral appeal, in which users are given more control over their actions, would perform even better.

However, we find that adding appeals (of any kind) does not significantly improve user perceptions of fairness compared to a no appeal baseline condition. In fact, while the differences are not significant, the no appeal condition performs better than all appeal types for all dependent

	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

[a]Values refer to the 7-point Likert scale rating, where a rating of 7 is most fair, accountable, trustworthy and in control.
[b]All results were controlled for participants' prior opinion about Facebook, prior experiences with account suspensions, and propensity to trust social media platforms.

Table 2. ANOVA Results.

measures: fairness, accountability, trustworthiness, and feelings of control. The results are shown in Table 2. Whether the appeal is evaluated by a person or by another algorithm does not significantly effect perceptions of FACT. A behavioral appeal, in which the user has time to change their behavior and address the potential suspension, has the lowest scores for all measures in comparison to the other conditions, though, again, the difference is small.

On the other hand, over the course of the experiment, perceptions of the fairness, trustworthiness and transparency of the process slightly increase (Table 3). In the case of perceptions of fairness and trustworthiness, these differences are significant. Although most participants did not describe explicit changes in their thinking, seeing the set of conflicting goals a platform needs to balance may have improved their perceptions of the fairness, as in our pilot tests.

One would expect current appeal designs to work well; Facebook has thousands of engineers to design effective appeal mechanisms. Our initial hypothesis was that the current appeal design would work better than none, and that a less punitive appeal would work even better. Instead, we find that none of these approaches improve user perceptions of fairness, accountability and trustworthiness. This is surprising and new. The lack of benefit for adding appeals is also surprising, given the responses users write — which often request appeals — as discussed in the next section.

Measures	Initial[a]	Final	M[b]	t	p[c]
Fair	2.87	3.25	0.38	t(181) = 3.31	0.001*
Unbiased	3.00	3.18	0.17	t(181) = 1.37	0.170
Trustworthy	2.92	3.30	0.37	t(181) = 3.07	0.002*
Transparent	3.08	3.19	0.10	t(181) = 0.89	0.373

[a]Values refer to the 7-point semantic scale rating, where a rating of 7 is most fair, trustworthy, etc.
[b]Mean of the differences
[c]*$p < 0.05$

Table 3. Paired t-test Results.

5.2 Writing Appeals

Our second research question addressed how users write appeals of a social media account suspension:

RQ2 What do users choose to contest in their appeals of algorithmic content moderation?

While past work has found that users most often attribute content moderation or suspensions to other users (who may have reported them maliciously) [76], we find many other arguments more common in our data, which often address algorithmic decision making processes more directly. We note that all participants were told that Facebook *incorrectly* identified them as spreading misinformation, so most arguments start with that assumption.

5.2.1 Contesting the Fundamental Goals. While many participants (n=61) mention their agreement with the basic goals (*if you do something wrong then you should have privileges taken away from you* [P27]), many others fundamentally argue that Facebook should not moderate content (n=48). Many of these participants refer to basic principles, like freedom of speech to argue that their content should not be removed (n=16): *Facebook has a right to remove offensive information but they don't have to right to suppress freedom of speech, including if it is different from their views or especially their political leanings* [P185]. And while many appreciate that Facebook was attempting to curb the spread of misinformation, others argue that Facebook should have other priorities than the spread of misinformation (n=11): *I think it's stupid. You need to focus more on the cyberbullying and catfishing* [P329] or that misinformation does not really harm other users: *If it isn't an illegal action activity taking place on the page, I think there's no harm to anyone nor Facebook* [P270].

Many users particularly respond to the intensity of the violation, noting that a suspension is so serious that it should not be a response for spreading misinformation (n=26): *To suspend someone for 'misinformation' is Big Brother at work. Suspensions should be the last straw for threatening someone, or vulgar language or porn* [P348]. Some even proposed alternative mechanisms that would avoid account suspension, for example: *I should be notified what is misinformation, that post should be deleted, but I should not have my account suspended* [P92].

5.2.2 Contesting Automation. A small number of participants agree that automation is a good approach (n=6) and simply suggest the system has encountered some kind of failure mode: *AI had a "brain fart" please recheck your system maybe reboot* [P328].

However, it was more common for users to contest the automation of the decision (n=125). Many users fundamentally disagree with automating account suspension decisions: *It was unfair. An algorithm should not determine final decisions* [P285], Many argue that automation cannot make these decisions, because it misinterprets or misses important context (n=18). For example, one participant notes how Facebook follows rules strictly and cannot adapt to different circumstances:

> *Facebook banned me for this very thing. Because they said the money in the board game [I was trying to sell] was counterfeit. It was Monopoly money, it's not supposed to be real money* [P115].

In some cases, participants even argue about exact automation techniques (n=22), often believing that text processing of some kind was responsible for the decision, and often arguing that text processing cannot account for tone, intent, and context:

> *A computer program is designed only to recognize words and their definitions and not the expressions behind them. It all has do to with how words are perceived not defined. A person can say something on a computer to someone and have it come across completely different than if said in person* [P170].

Another describes crucial contextual information – the user describes a post that *was* in fact misinformation, but where they were critiquing it, not simply sharing it further:

> *You need to use something other than a computer program to decide whether a person is actually guilty of spreading misinformation. I doubt your automated system can understand the context of a person's post. If you had a human reviewing these things then they would have seen that I was not spreading any misinformation, but merely reposting and then responding to it. I request that an actual human being look over the alleged misinformation and then make a judgement* [P340].

This request to have a human review the decision was extremely common (n=51). Indeed, while the appeals conditions appeared to have little impact on FACT perceptions, participants often asked for an appeal directly (n=33). Participants were well aware and seemed quite understanding that any system can make mistakes, but when those mistakes occur they believe *there should be double checks to ensure accounts aren't just randomly closed* [P282].

5.2.3 Contesting the Opacity of the System. Many participants asked for explanations and further details about why their account had been suspended (n=113). Some users write that they simply don't understand (n=22): *I have no idea what you mean; I don't spread false news* [P389], while others ask for the reasons (n=45): *I would say that I have no idea why they are accusing me of this, and request to know the reasons* [P26]. Some participants write that they believe they know the reasons that Facebook uses, but find them difficult to understand and interpret (n=22): *Policies are often vague or not written in simple English* [P235].

In many cases, participants ask for the exact post that caused the suspension (n=27), often writing that this would help them provide a better appeal by providing context: *What in particular is the article or post that caused this issue so I can further elaborate my reasons and give you an explanation on the article* [P399], or that the system should provide more transparency and specificity into the decision making process:

> *Unless you are on the inside, there is no way the average user can determine the algorithms that are used to provide data, posts and information. Also, the statement that was posted was vague without making any specific references as to the cause for the suspension* [P112].

In addition, as we will discuss below, many users want to know how other posts and accounts are treated, in order to understand how they should argue their case.

5.2.4 Contesting Inconsistency. Many participants argue that Facebook's approach is inconsistent (n=84). Often participants believe the system is biased (n=59): *Facebook, your bias is showing. what "misinformation" has been attributed to me?* [P324] or supporting some unknown, undesirable goals *Facebook has secret agendas* [P257] – arguments which often co-occur with expressions of anger (n=25): *You're an idiot so unlock my account* [P297] or hatred of Facebook (n=13): *I personally think Facebook is the freaking devil* [P200]. This assumption of bias is particularly common for conservatives who frequently believe their accounts are disproportionately targeted (n=29): *Facebook suspends too many conservative accounts and very few liberal ones* [P171].

Sometimes they argue that Facebook is not following its own policies: *Facebook policy was not followed* [P290] or is inconsistent in how it applies rules: *I think they suck. They will suspend one person and another who does the same thing is still using Facebook* [P105]. Several users refer to past experiences with Facebook and perceived inconsistencies in how it applies rules:

> *I have friends who were producing a series of photographs portraying stereotypes of African Americans, but making the stereotype a strength or beauty, not a negative. This series included a mother nursing a baby, surrounded by her other children. It was a beautiful photo of a strong mother figure, but Facebook took it down. Kim Kardashian had her*

> *naked tush plastered across Facebook, but a professional photographer couldn't have his sister nursing her baby????? Unfair, and just wrong* [P315].

As this suggests, many participants are concerned about whether they were treated similarly to other users: *I am not spreading misinformation mainly because the majority of my posts are shared from other people. If me sharing things from other people qualifies as misinformation, everyone should have their accounts suspended* [P57]. Indeed, in contesting this inconsistency, users often called for explanations or for examples to compare with other users: *It is inconsistent, arbitrary and never, ever explained properly or understandably to the User. They do what they want when they want* [P348].

5.2.5 Contesting the Decision. Users also contest the decision in three major ways:
The Data Defends Itself Many users argue that the error is obvious (*Robots made error* [P311]) from the data and information in their account (n=84). In many cases the participants argue that it is self-evident that the type of content they put on Facebook would never be considered misinformation (n=21): *the information they recieved is incorrect considering I only post family pics* [P192] or *I only post funny pictures and jokes. They can clearly see* [P372]. And many note that they rarely post anything, so there simply is no content to be considered misinformation (n=28): *I rarely post anything at all, so I don't see how my account could possibly be spreading misinformation* [P198]. Sometimes participants argue that the rules are fundamentally wrong and that a certain type of content should not be considered misinformation (n=6): *I am simply sharing my opinion on certain topics; All of my posts are just my personal opinion and what I believe to be true* [P66]. This suggests that opinions could not be misinformation, only claims about facts could; and in general that Facebook need only identify content as opinion (or funny or family related) and then would know that it could not be misinformation.

Adding Context Despite the fact that all participants are told they are *incorrectly* identified as sharing misinformation, many accept that the content they posted may have been interpreted as misinformation. Some believe their content was identified as misinformation, but wish to provide context either about the post itself or about their account (n=84). A few even agree with the algorithmic decision: *With all information available. Maybe they are seeing something I am not aware of* [P363].

Some wish to defend the posts, frequently by arguing that they can provide sources and references to support what they posted (n=14):

> *The information that I provide through my account has been deeply researched and is based only on written facts. Facts on which I'm more than willing to share with your department* [P359].

Others suggest that they did post the content described, but there was context or some other reason for posting it. Participants provide additional context by describing, for example, why they were commenting on a misinformation post: *sometimes I needed to listen someone's opinion* [P50].

Others provide context about themselves. Many participants say that they might've actually posted misinformation, but their intentions should be taken into account more than the actual post (n=45). Many of these participants noted that it was an unintended accident that they would correct in the future: *unaware of misuse of account and if suspension is lifted will use caution in future posts* [P241]. Indeed, many participants write that they believed the information themselves, *I got the information through Facebook so I posted thinking that it was correct info*, and thought that Facebook ought to help them to correct it: *if I'm posting things that are not true I would want to know about them and get them right* [P139]. Further, many participants made arguments in the virtue ethics style, arguing that they were not the kind of person who would share misinformation (n=23):

I have never knowingly misled anyone in any way [P14] and *I am honest and not the type of person who would do this at all* [P193].

Outside Forces In addition, many users provide alternative explanations for how misinformation was shared from their account. These typically attribute the posting or identification of misinformation to some external forces at work (n=54). One of the most popular was arguing that their account had been hacked (n=36): *Maybe I've been hacked by someone or something. I just don't know, but I'd like to keep my account with you guys* [P280].

Finally, as has been found in prior work [76], many users think that a person was responsible for reporting their content even though it was not actually misinformation (n=18):

> *I know what I post and it's not about any foolishness, [...] evidently some one evil has played some kind of joke of which I am not aware of and know nothing about* [P242].

In these cases, the external force was another user of the system, like *a former friend wanted to start something* [P175], and Facebook would need to understand these social factors to ignore a malicious report.

5.2.6 Leaving the system. Many users say that they would leave the system if they were incorrectly identified as sharing misinformation (n=58). While some users said they would ignore the notification because they think it is a scam (n=6, *I assume this is a scam* [P285]), many more said they would ignore it because they don't care about their account being closed (n=17): *I do not care if you close my account* [P237]. Indeed, many participants said they would not bother appealing and would themselves close their account *I wouldn't explain, I'd leave facebook...* [P347] or *Dear Mark Zuckerberg, Goodbye* [P315]. Unfortunately some participants chose this path because they felt fatalistic about the decision and that they were unlikely to influence the final decision (n=15): *I would get off Facebook because it's harder to prove its not true than it is to not try to fight it* [P234]. Indeed many participants report that *at the end of the day Facebook does what it wants* [P106] (n=27), often suggesting that users don't influence its decisions. This kind of fatalism or resignation has been found in the face of other algorithmic governance and decision making systems [100]. Since a feeling of self-efficacy and control is essential for human well-being [66], we had hoped that appeals would increase the perception of feelings of control. As we discuss below, it may be that the current design of appeals is not able to address this need.

6 LIMITATIONS

The experimental design used avoids the self-selection bias of some prior work; however, this work may showcase one of the disadvantages of a representative sample. While we believe most of the lack of effect should be attributed to poor appeal design, stronger effects might be found among communities that have experienced marginalization and negative effects at the hands of social media platforms in the past (e.g., queer users [15, 22], people of color [2, 13, 43], and so on).

The experiment also used hypothetical scenarios, which while very common in research on perceived fairness, does mean that we cannot be certain that people's written explanations and plans were what they would actually write to Facebook. In combination with previous research on self-selected reports [76], this project helps converge on users' experience of contesting algorithmic content moderation. However, future work might augment these approaches using proprietary data from user appeals provided by Facebook or other social media, or with in-lab experiments.

Finally, we restricted our study to consider only Facebook. Recent polling has found little public confidence in social media companies' abilities in general when conducting content moderation [61], but Facebook in particular has faced extensive public backlash. For example, recent work focused on social media platforms as a source of political news found that more people distrust Facebook (59%)

compared to all other social media platforms (Twitter 48%, Instagram 42%, Youtube 36%, Reddit 24%) [52]. Facebook has also faced significant criticism (and popular press coverage) for its policy decisions, from the early outcry over its research policy [41] to more recent fact checking [63], ad placement [14], and privacy policies [96]. While it is possible these findings will generalize to other platforms, more work is needed. And given the interplay between technical design and policy decisions in user perceptions, this greater level of distrust and heightened attention may make improving perceptions for Facebook more challenging than other platforms.

7 DISCUSSION

While this research focused on the potential benefits of appeals, they also have systemic failures. We outline several, before offering design improvements suggested by our work.

7.1 Systematic Failures of Appeals

7.1.1 Power Imbalances. For decades — and continuing today — many have expressed concern about the role of money in the justice system [32, 34, 47], "*As Justice Black has stated, 'There can be no equal justice where the kind of trial a man gets depends on the amount of money he has'* [34]. And while Forer's suggestion that the wealthy can avail themselves of "*endless appeals*" may not be literally true, it is evocative of a reality where cases can often be won by refusing to give up — and continuing to spend money on appeals that other litigants do not have [32]. Designing for appeals in sociotechnical systems face similar issues of asymmetries of wealth, power, and access.

7.1.2 Decisions in Isolation. As many researchers have noted, current decisions are highly flexible (or inconsistent) [76], and as our work and others have found [22], this perceived inconsistency is one of the major concerns users express with the platform. Fundamentally, appeal systems are designed to handle decisions individually. So when improvements have been made to protect minority interests (for example, updates to gender categorization) they have been ad hoc [20, 37]. This makes any systematic improvements to social media platforms highly dependent on how organized communities are, how powerful they are, and how much access they have to platform-specific decision makers. While Facebook has suggested that a "social media supreme court" should adjudicate such decisions in the past [77], little progress has been made. Some have argued that this is a means for Facebook to avoid accountability, but it might provide a more equitable, open, transparent process for changing social media platform decisions.

7.1.3 Perceived Illegitimacy. Prior research showed the impact of voice on perceived fairness and satisfaction with decision making systems [68]. However, experimental work has also found that "*voice-based procedural justice effects can be attenuated or reversed when the procedure is weak or when there is social criticism of the procedure*" [68]. One explanation for our limited effect may be that the current design has faced social criticism – particularly of bias against conservative users. It may be the case that, in the face of broad social criticism of social media platforms as a whole, that appeals on these platforms can never provide users the voice they need.

7.2 Better Design of Appeals

Nevertheless, our work does indicate some directions for better design of appeal for social media systems. We highlight a few possible directions.

7.2.1 Scaffolding the Appeal. The experiment tested appeals where users provide unstructured explanations. But this may not be the optimal design for appeals. Researchers have shown that providing scaffolding can help users improve their learning [3], particularly for complex tasks like writing [88]. Our review of existing appeal systems found some provide such scaffolds [17].

Scaffolding helps learning in part by providing a model of the appropriate structure, soliciting information that is needed [3]. An appeal system could do so by providing a structured form, suggesting what information or elements of an argument are considered acceptable or relevant, as in the credit scoring appeals. Our work found that many people wished to provide additional contextual information; platforms could identify which kinds of context are considered in decisions. For example, Facebook policies on hate speech include exemptions like "*Sometimes people share content containing someone else's hate speech for the purpose of raising awareness or educating others*" [29]. The appeal system might lay out for users which kinds of context, like responding to raise awareness, are considered in the decision making. And in doing so, can help address the perceived inconsistency of the decision making process.

In addition, research on crowdwork writing tasks found that even relatively low cost scaffolding mechanisms (like highlighting) can be an effective way to communicate areas that need work [88]. Reviewers might quickly highlight parts of the appeal that were not persuasive to focus users' attention and build mutual understanding.

7.2.2 Unburdening the User. Existing appeals processes shift many burdens from the platform to the user: emotional regulation, understanding complex documentation, formulating effective arguments. As our results show, users struggle with many aspects of this, as with participants who expressed anger when trying to construct their appeals: "*You're an idiot so unlock my account*" [P297]. While this anger is unlikely to be persuasive to platforms, it is difficult for users to manage all of these tasks while being faced with an account suspension that could harm their relationships, businesses, and ability to access other services. Borrowing strategies from other formal and informal appeal systems could help shift these burdens back to the platform. For example, platforms could build systems that allow users to directly express their anger and frustration. One prototype technology facilitated somatic experience, by storing inappropriate screams for later [21]. Social media appeal systems could support expressive, experiential equivalents.

Being conscious of the power imbalance that exists with end users, and ensuring that appeals are not costly in a way that prioritizes some users at the expense of others, would also ensure more equitable systems. Instead of prioritizing users who understand the systems better or make more experienced arguments, systems could allow synchronous conversations with decision makers (or automated agents) to help users better understand the decision making and make their case. Platforms might even borrow the approach used in the legal system by allowing professional agents to present an argument on the users behalf. Finally, platforms could organize communities within appeal systems, perhaps in tandem with auditing visualizations or tools, to ensure that groups with more existing access to decision makers or the press would not have preferential treatment. All of these approaches would relieve users of some of their current burdens.

7.2.3 Constructing the User. Few users read social media platforms "educational materials," like community standards and terms of service. But if users experience content moderation, or an account suspension, an appeal process offers an opportunity – instead of facing punishment – to develop the decision subject into a better social media user. Integrating the Terms of Service and Community Standards on Facebook closer to the appeal location could also help. However, as some of our participants noted, as written now, these policies can be difficult to read and interpret: *Policies are often vague or not written in simple English* [P235]. Using alternative strategies to communicate policies, like videos, stories, or examples, could be an important aspect of addressing perceived legitimacy of the platform. And more closely linking the appeal system to the documentation could help users understand how decisions are made systematically.

As our results found, many users make appeals to "virtue ethics" style arguments, emphasizing that they are not the kind of person who would spread misinformation, as with those who wrote: *if*

I'm posting things that are not true I would want to know about them and get them right [P139] or *I am honest and not the type of person who would do this at all* [P193]. By educating users in the platform about what misinformation is and having them reflect on how they could avoid spreading it, platforms could leverage these good intentions to make inroads in challenging problems like the growth of 'fake news.'

7.3 Moving Towards Contestability

Content moderation decisions make fundamental choices about what speech should be supported or silenced. Many have argued that social media has become an important "public square," despite private ownership [51, 92]. Our results showed that many users have concerns about "free speech" on these platforms. And if social media platforms constitute a public square, then the public has an interest in ensuring that important speech is not being silenced.

However, the tacit assumption of many content moderation systems is that users are often malicious and use any information provided to game the system. Theoretical work focuses on formal models of accountability, with an emphasis on account suspension as a form of punishment [30]. Malicious users exist and do need punishment [72], but this approach harms well meaning users [15].

We suggest that users, platforms, and regulators need to fundamentally re-evaluate the purpose of content moderation and the appeal processes for content moderation. Is it only to change the outcome for a particular post? Or can appeals be used as a mechanism for user feedback, to let the platform know that it is systematically erring in its approach? A framing focused on punishment eliminates the potential for content moderation to serve educational functions [76], and, in the tradition of human-computer interaction, to allow users — *and the system* — to recover from error.

Platforms have the opportunity to build such true contestability into their systems through appeal designs; and in doing so, they can address the challenges of perceived illegitimacy, inconsistency and power asymmetries. But to do so, they will need to embrace an approach that takes a user-focus in all aspects: design, code, but also policy. Appeals could integrate information about how decisions are made, so users can learn, reflect on, and share whether they agree with those decisions. The platform could not only audit its own appeals to improve decision making, but surface appeals to users to benefit individuals (e.g., showing whether they were treated consistently to others) and communities (e.g., providing an opportunity to discover and collectively advocate for changes) [73, 76, 89]. In this case, appeals shift from a linear to an iterative process, in which on-going decisions are negotiated in dialogue with end users and communities. While this will require substantial changes from the platforms — allowing user input into policy, infrastructure, and design — this work begins to showcase how contestability might extend beyond experts [42, 56, 57] to work for all.

8 CONCLUSION

In this work, we test different types of appeals to see how they impact users' experiences of contesting automated content moderation. We explore appeals very similar to Facebook's current design, but also alternatives designed to address some of the challenges of content moderation systems: algorithmically-reviewed and behavioral appeals. However, none improve FACT perceptions. This suggests that these designs have not yet achieved the benefits that prior work has identified for having one's voice be heard; future work should seek to identify better designs.

We also qualitatively analyze how users write appeals, and find that they contest the decision itself, but also more fundamental issues like the goal of moderating content, the idea of automation, and the inconsistency of the system as a whole. However, the strength of users' responses suggest the opportunity to use appeal systems to allow users to co-construct the decision making process and the importance of doing so – when users feel they cannot influence a decision, they express fatalism and desire to abandon the system.

REFERENCES

[1] Edgar Alvarez. 2019. Instagram will soon let you appeal post takedowns. *Engadget* (2019). https://www.engadget.com/2019/05/07/instagram-appeals-content-review-taken-down-posts/

[2] Julia Angwin and Hannes Grassegger. 2017. Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children. https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms. ProPublica.

[3] Arthur N. Applebee and Judith A. Langer. 1983. Instructional Scaffolding: Reading and Writing as Natural Language Activities. *Language Arts* 60, 2 (1983), 168–175.

[4] American Bar Association. 2017. How Courts Work: Steps in a Trial. https://www.americanbar.org/groups/public_education/resources/law_related_education_network/how_courts_work/appeals.html.

[5] Lawrence C Becker and Charlotte B Becker. 2013. *Encyclopedia of ethics*. Routledge.

[6] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proc. CHI*. ACM. https://doi.org/10.1145/3173574.3173951

[7] Steven L Blader and Tom R Tyler. 2003. A four-component model of procedural justice: Defining the meaning of a "fair" process. *Personality and Social Psychology Bulletin* 29, 6 (2003), 747–758.

[8] John Braithwaite. 2002. *Restorative justice & responsive regulation*. Oxford University Press.

[9] Joel Brockner, Phyllis A Siegel, Joseph P Daly, Tom Tyler, and Christopher Martin. 1997. When trust matters: The moderating effect of outcome favorability. *Administrative science quarterly* (1997), 558–583.

[10] Joel Brockner and Batia Wiesenfeld. 2005. How, when, and why does outcome favorability interact with procedural fairness? In *Handbook of organizational justice*, J. Greenberg and J. A. Colquitt (Eds.). Lawrence Erlbaum Associates Publishers.

[11] Emile G Bruneau and Rebecca Saxe. 2012. The power of being heard: The benefits of 'perspective-giving' in the context of intergroup conflict. *Journal of experimental social psychology* 48, 4 (2012), 855–866.

[12] James Carson. 2019. Fake news: What exactly is it ? and how can you spot it? https://www.telegraph.co.uk/technology/0/fake-news-exactly-has-really-had-influence/. The Telegraph.

[13] Online Censorship. [n.d.]. OFFLINE-ONLINE: data visuals exploring how content moderation practices impact marginalized communities. https://onlinecensorship.org/content/infographics.

[14] Bill Chappell. 2020. FEC Commissioner Rips Facebook Over Political Ad Policy: 'This Will Not Do'. https://www.npr.org/2020/01/09/794911246/fec-commissioner-rips-facebook-over-political-ad-policy-this-will-not-do.

[15] Alexander Cheves. 2018. The Dangerous Trend of LGBTQ Censorship on the Internet. *Out Magazine* (2018). http://www.out.com/out-exclusives/2018/12/06/dangerous-trend-lgbtq-censorship-internet

[16] Jane Coaston. 2019. The Facebook free speech battle with conservatives, explained. https://www.vox.com/technology/2019/5/6/18528250/facebook-speech-conservatives-trump-platform-publisher. Vox.

[17] Federal Trade Commission. 2013. Sample Letter for Disputing Errors on Your Credit Report. https://www.consumer.ftc.gov/articles/0384-sample-letter-disputing-errors-your-credit-report.

[18] Federal Trade Commission. 2017. Disputing Errors on Credit Reports. https://www.consumer.ftc.gov/articles/0151-disputing-errors-credit-reports.

[19] Kate Crawford and Tarleton Gillespie. 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society* 18, 3 (2016), 410–428.

[20] Facebook Diversity. 2015. Last year we were proud to add... https://www.facebook.com/facebookdiversity/posts/last-year-we-were-proud-to-add-a-custom-gender-option-to-help-people-better-expr/774221582674346/.

[21] Kelly Dobson. 2005. Wearable Body Organs: Critical Cognition Becomes (Again) Somatic. In *Proc. Creativity & Cognition*. https://doi.org/10.1145/1056224.1056267

[22] Stefanie Duguay, Jean Burgess, and Nicolas Suzor. 2018. Queer womens experiences of patchwork platform governance on Tinder, Instagram, and Vine. *Convergence: The International Journal of Research into New Media Technologies* (2018). https://doi.org/10.1177/1354856518781530

[23] Elizabeth Dwoskin, Jeanne Whalen, and Regine Cabato. 2019. Content moderators at YouTube, Facebook and Twitter see the worst of the web and suffer silently. https://www.washingtonpost.com/technology/2019/07/25/social-media-companies-are-outsourcing-their-dirty-work-philippines-generation-workers-is-paying-price. The Washington Post.

[24] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, A Elazari, Eric Gilbert, and Karrie Karahalios. 2019. User Attitudes towards Algorithmic Opacity and Transparency in Online Reviewing Platforms. In *Proc. CHI*.

[25] Facebook. [n.d.]. Disabled Accounts. Facebook Help Center. https://www.facebook.com/help/185747581553788/

[26] Facebook. [n.d.]. My Personal Account Was Disabled. Facebook Help Center. https://www.facebook.com/help/contact/260749603972907

[27] Facebook. 2018. Enforcing Our Community Standards. Facebook Newsroom. https://newsroom.fb.com/news/2018/08/enforcing-our-community-standards/
[28] Facebook. 2019. Facebook Reports Third Quarter 2019 Results. https://investor.fb.com/investor-news/press-release-details/2019/Facebook-Reports-Third-Quarter-2019-Results/default.aspx.
[29] Facebook. 2020. Community Standards: Hate Speech. https://www.facebook.com/communitystandards/hate_speech.
[30] Joan Feigenbaum, Aaron D. Jaggard, and Rebecca N. Wright. 2011. Towards a Formal Model of Accountability. In *Proc. New Security Paradigms Workshop*. ACM. https://doi.org/10.1145/2073276.2073282
[31] Jessica K Flake, Jolynn Pek, and Eric Hehman. 2017. Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science* (2017), 370–378.
[32] Lois G Forer. 1984. *Money and Justice: Who Owns the Courts?* Norton.
[33] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press.
[34] Arthur Goldberg. 1964. Equal Justice for the Poor, Too; Far too often, money–or the lack of it–can be the deciding factor in the courtroom, says Justice Goldberg, who calls for a program to insure justice for all Americans. *New York Times* (March 1964).
[35] Bryce Goodman and Seth Flaxman. 2016. European Union regulations on algorithmic decision-making and a "right to explanation". *arXiv preprint arXiv:1606.08813* (2016).
[36] John Gramlich. 2019. 10 facts about Americans and Facebook. https://www.pewresearch.org/fact-tank/2019/05/16/facts-about-americans-and-facebook/. Pew Research Center.
[37] Brandon Griggs. 2014. Facebook goes beyond 'male' and 'female' with new gender options. https://www.cnn.com/2014/02/13/tech/social-media/facebook-gender-custom/index.html. *CNN* (2014).
[38] Angela T Hall. 2005. *Accountability in organizations: An examination of antecedents and consequences.* Ph.D. Dissertation. Florida State University.
[39] Jon Henley. 2018. Global crackdown on fake news raises censorship concerns. https://www.theguardian.com/media/2018/apr/24/global-crackdown-on-fake-news-raises-censorship-concerns. The Guardian.
[40] David A Hensher. 2010. Hypothetical bias, choice experiments and willingness to pay. *Transportation Research Part B: Methodological* 44, 6 (2010), 735–752.
[41] Alex Hern. 2014. Facebook T&Cs introduced 'research' policy months after emotion study. https://www.theguardian.com/technology/2014/jul/01/facebook-data-policy-research-emotion-study.
[42] Tad Hirsch, Kritzia Merced, Shrikanth Narayanan, Zac E Imel, and David C Atkins. 2017. Designing contestability: Interaction design, machine learning, and mental health. In *Proc. DIS*. ACM, 95–99.
[43] Amanda Holpuch. 2015. Facebook still suspending Native Americans over 'real name' policy. *The Guardian* (February 2015). https://www.theguardian.com/technology/2015/feb/16/facebook-real-name-policy-suspends-native-americans
[44] John W Huppertz, Sidney J Arenson, and Richard H Evans. 1978. An application of equity theory to buyer-seller exchange situations. *Journal of marketing research* 15, 2 (1978), 250–260.
[45] Mikella Hurley and Julius Adebayo. 2016. Credit scoring in the era of big data. *Yale Journal of Law & Technology* 18 (2016), 148.
[46] Ben Hutchinson and Margaret Mitchell. 2019. 50 Years of Test (Un) fairness: Lessons for Machine Learning. In *Proc. FAT**. https://doi.org/10.1145/3287560.3287600
[47] Liane Jackson. 2019. Balance of power: Money and inequity in the judicial system. http://www.abajournal.com/magazine/article/balance-of-power.
[48] Tracy Jan and Elizabeth Dwoskin. 2017. A white man called her kids the n-word. Facebook stopped her from sharing it. 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. The Washington Post.
[49] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2018. Does Transparency in Moderation Really Matter?: User Behavior After Content Removal Explanations on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 2 (2018).
[50] Thomas M Jones. 1991. Ethical decision making by individuals in organizations: An issue-contingent model. *Academy of management review* 16, 2 (1991), 366–395.
[51] Jeffrey S Juris. 2012. Reflections on# Occupy Everywhere: Social media, public space, and emerging logics of aggregation. *American Ethnologist* 39, 2 (2012), 259–279.
[52] Mark Jurkowitz and Amy Mitchell. 2020. An oasis of bipartisanship: Republicans and Democrats distrust social media sites for political and election news. *Pew Research Center* (January 2020).
[53] Natascha Just and Michael Latzer. 2017. Governance by algorithms: reality construction by algorithmic selection on the Internet. *Media, Culture & Society* 39, 2 (2017).

[54] Sheryl E Kimes. 1994. Perceived Fairness of Yield Management: Applying yield-management principles to rate structures is complicated by what consumers perceive as unfair practices. *Cornell Hotel and Restaurant Administration Quarterly* 35, 1 (1994), 22–29.
[55] Kate Klonick. 2017. The new governors: The people, rules, and processes governing online speech. *Harvard Law Review* 131 (2017), 1598.
[56] Daniel Kluttz, Nitin Kohli, and Deirdre K Mulligan. 2018. Contestability and Professionals: From Explanations to Engagement with Algorithmic Systems. *Available at SSRN 3311894* (2018).
[57] Daniel Kluttz and Deirdre K Mulligan. 2019. Automated decision support technologies and the Legal Profession. *Berkeley Technology Law Journal* (2019).
[58] Jason Koebler and Joseph Cox. 2018. The Impossible Job: Inside Facebook?s Struggle to Moderate Two Billion People. (August 2018). https://www.vice.com/en_us/article/xwk9zd/how-facebook-content-moderation-works.
[59] Laura J Kray and E Allan Lind. 2002. The injustices of others: Social reports and the integration of others' experiences in organizational justice judgments. *Organizational behavior and human decision processes* 89, 1 (2002), 906–924.
[60] Peter Kuppens, Iven Van Mechelen, Dirk JM Smits, Paul De Boeck, and Eva Ceulemans. 2007. Individual differences in patterns of appraisal and anger experience. *Cognition and Emotion* 21, 4 (2007), 689–713.
[61] John Laloggia. 2019. U.S. Public has little confidence in social media companies to determine offensive content. *Pew Research Center* (July 2019).
[62] Kyle Langvardt. 2017. Regulating Online Content Moderation. *Georgetown Law Journal* 106 (2017), 1353.
[63] Dave Lee. 2019. Facebook's Zuckerberg grilled over ad fact-checking policy. https://www.bbc.com/news/technology-50152062.
[64] Min Kyung Lee and Su Baykal. 2017. Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division.. In *Proc. CSCW*. 1035–1048.
[65] Min Kyung Lee, Ji Tae Kim, and Leah Lizarondo. 2017. A human-centered approach to algorithmic services: Considerations for fair and motivating smart community service management that allocates donations to non-profit organizations. In *Proc. CHI*. ACM.
[66] Lauren A Leotti, Sheena S Iyengar, and Kevin N Ochsner. 2010. Born to choose: The origins and value of the need for control. *Trends in cognitive sciences* 14, 10 (2010), 457–463.
[67] Gerald S Leventhal. 1976. What Should Be Done with Equity Theory? New Approaches to the Study of Fairness in Social Relationships. (1976).
[68] E Allan Lind and Tom R Tyler. 1988. *The social psychology of procedural justice*. Springer Science & Business Media.
[69] Amber Madison. 2015. When Social-Media Companies Censor Sex Education. https://www.theatlantic.com/health/archive/2015/03/when-social-media-censors-sex-education/385576/. The Atlantic.
[70] Nace Magner, Robert B Welker, and Terry L Campbell. 1995. The interactive effect of budgetary participation and budget favorability on attitudes toward budgetary decision makers: a research note. *Accounting, Organizations and Society* 20, 7-8 (1995), 611–618.
[71] Aaron Mak. 2018. Facebook content moderation rules: How company decides what to remove. https://slate.com/technology/2018/04/facebook-content-moderation-rules-how-company-decides-what-to-remove.html. Slate.
[72] Ariadna Matamoros-Fernandez. 2017. Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. *Information, Communication & Society* 20, 6 (2017), 930–946.
[73] J Matias, Amy Johnson, Whitney Erin Boesel, Brian Keegan, Jaclyn Friedman, and Charlie DeTar. 2015. Reporting, reviewing, and responding to harassment on Twitter. *Available at SSRN 2602018* (2015).
[74] Neal P Mero, Rebecca M Guidice, and Steve Werner. 2014. A field study of the antecedents and performance consequences of perceived accountability. *Journal of Management* 40, 6 (2014), 1627–1652.
[75] Adam Mosseri. 2017. Working to Stop Misinformation and False News. Facebook Blog. https://www.facebook.com/facebookmedia/blog/working-to-stop-misinformation-and-false-news
[76] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* 20, 11 (2018), 4366–4383.
[77] Casey Newton. 2018. Facebook wants a social media supreme court so it can avoid hard questions. https://www.theverge.com/2018/4/3/17191198/facebook-supreme-court-mark-zuckerberg-content-moderation. *The Verge* (2018).
[78] Department of Health and Human Services Centers for Medicare & Medicaid Services. [n.d.]. *Medicare Redetermination Request Form – 1st Level of Appeal*. Form CMS-20027 (12/10).
[79] Sarah Perez. 2019. Twitter now lets users appeal violations within its app. *Tech Crunch* (2019). https://techcrunch.com/2019/04/02/twitter-now-lets-users-appeal-violations-within-its-app/
[80] Lewis Petrinovich, Patricia O'Neill, and Matthew Jorgensen. 1993. An empirical study of moral intuitions: Toward an evolutionary ethics. *Journal of personality and social psychology* 64, 3 (03 1993), 467–478.
[81] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *Proc. CHI*. ACM.

[82] R Eric Reidenbach and Donald P Robin. 1990. Toward the development of a multidimensional scale for improving evaluations of business ethics. *Journal of business ethics* 9, 8 (1990), 639–653.
[83] Maria A. Ressa. 2018. Facebook suspends account of Pinoy Ako Blog's Jover Laurio. *Rappler* (2018). https://www.rappler.com/technology/social-media/203269-jover-laurio-facebook-account-suspended Updated 7:45 PM, May 25, 2018.
[84] Sarah Roberts. 2016. Commercial Content Moderation: Digital Laborers' Dirty Work. In *The Intersectional Internet: Race, Sex, Class and Culture Online*, S. U. Noble and B. Tynes (Eds.). Peter Lang Publishing.
[85] Sarah Roberts. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media.* Yale University Press.
[86] Sarah T Roberts. 2018. Digital detritus: 'Error' and the logic of opacity in social media content moderation. *First Monday* 23, 3 (2018).
[87] Julian B Rotter. 1967. A new scale for the measurement of interpersonal trust 1. *Journal of personality* 35, 4 (1967), 651–665.
[88] Niloufar Salehi, Jaime Teevan, Shamsi Iqbal, and Ece Kamar. 2017. Communicating context to the crowd for complex writing tasks. In *Proc. CSCW*.
[89] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *ICA Preconference: Data and discrimination: converting critical concerns into productive inquiry* (2014).
[90] Nripsuta Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David Parkes, and Yang Liu. 2018. How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness. In *Proc. AIES*.
[91] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* (2019).
[92] Bill Sherman. 2011. Your Mayor, Your Friend: Public Officials, Social Networking, and Unmapped New Public Square. *Pace Law Review* 31 (2011), 95.
[93] Ellen Silver. 2018. Hard Questions: Who Reviews Objectionable Content on Facebook? And Is the Company Doing Enough to Support Them? Facebook Newsroom. https://newsroom.fb.com/news/2018/07/hard-questions-content-reviewers/
[94] Olivia Solon. 2017. Facebook is hiring moderators. But is the job too gruesome to handle? https://www.theguardian.com/technology/2017/may/04/facebook-content-moderators-ptsd-psychological-dangers. The Guardian.
[95] Olivia Solon. 2017. Underpaid and overburdened: the life of a Facebook moderator. https://www.theguardian.com/news/2017/may/25/facebook-moderator-underpaid-overburdened-extreme-content. *The Guardian* (May 2017).
[96] Emily Stewart. 2018. Mark Zuckerberg testimony: the Facebook data privacy question he won?t answer. https://www.vox.com/policy-and-politics/2018/4/11/17225518/mark-zuckerberg-testimony-facebook-privacy-settings-sharing.
[97] Jeroen Stouten, David De Cremer, and Eric Van Dijk. 2006. Violating equality in social dilemmas: Emotional and retributive reactions as a function of trust, attribution, and honesty. *Personality and Social Psychology Bulletin* 32, 7 (2006), 894–906.
[98] Jeroen Stouten, Peter Kuppens, and Stijn Decoster. 2013. Being angry for different reasons: The role of personality in distributive justice. *Journal of Applied Social Psychology* 43, 4 (2013), 795–805.
[99] Nicolas Suzor, Tess Van Geelen, and Sarah Myers West. 2018. Evaluating the legitimacy of platform governance: A review of research and a shared research agenda. *International Communication Gazette* 80, 4 (2018).
[100] Joseph Turow, Michael Hennessy, and Nora Draper. 2015. The tradeoff fallacy: How marketers are misrepresenting American consumers and opening them up to exploitation. *Available at SSRN 2820060* (2015).
[101] Tom R Tyler, Kenneth A Rasinski, and Nancy Spodick. 1985. Influence of voice on satisfaction with leaders: Exploring the meaning of process control. *Journal of Personality and Social Psychology* 48, 1 (1985), 72.
[102] Kristen Vaccaro, Dylan Huang, Motahhare Eslami, Christian Sandvig, Kevin Hamilton, and Karrie Karahalios. 2018. The Illusion of Control: Placebo Effects of Control Settings. In *Proc. CHI*. https://doi.org/10.1145/3173574.3173590
[103] Kristen Vaccaro, Karrie Karahalios, Deirdre Mulligan, Daniel Kluttz, and Tad Hirsch. 2019. Contestability in Algorithmic Decision Making. In *Proc. CSCW Workshops*. ACM.
[104] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proc. CHI*. ACM. https://doi.org/10.1145/3173574.3174230
[105] Ian Wren. 2018. Facebook Updates Community Standards, Expands Appeals Process. *NPR* (2018). https://www.npr.org/2018/04/24/605107093/facebook-updates-community-standards-expands-appeals-process
[106] Jillian C. York and Corynne McSherry. 2019. Content Moderation is Broken. Let Us Count the Ways. | Electronic Frontier Foundation. https://www.eff.org/deeplinks/2019/04/content-moderation-broken-let-us-count-ways. Electronic Frontier Foundation.

Received January 2020; revised June 2020; accepted July 2020