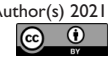
Article

new media & society
1–23



DOI: 10.1177/14614448211032310
journals.sagepub.com/home/nms


Common sense or censorship: How algorithmic moderators and message type influence perceptions of online content deletion

João Gonçalves
Erasmus University, The Netherlands

Ina Weber
Erasmus University, The Netherlands; University of Antwerp, Belgium

Gina M. Masullo
University of Texas at Austin, USA

Marisa Torres da Silva
Universidade Nova de Lisboa, Portugal

Joep Hofhuis
Erasmus University, The Netherlands

Abstract

Hateful content online is a concern for social media platforms, policymakers, and the public. This has led high-profile content platforms, such as Facebook, to adopt algorithmic content-moderation systems; however, the impact of algorithmic moderation on user perceptions is unclear. We experimentally test the extent to which the type of content being removed (profanity vs hate speech) and the explanation given for its removal (no explanation vs link to community guidelines vs specific explanation)

Corresponding author:
João Gonçalves, Department of Media & Communication, Erasmus University, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands.
Email: ferreiragoncalves@eshcc.eur.nl

influence user perceptions of human and algorithmic moderators. Our preregistered study encompasses representative samples ($N = 2870$) from the United States, the Netherlands, and Portugal. Contrary to expectations, our findings suggest that algorithmic moderation is perceived as more transparent than human, especially when no explanation is given for content removal. In addition, sending users to community guidelines for further information on content deletion has negative effects on outcome fairness and trust.

Keywords
Artificial intelligence, content moderation, cross-country, experiment, hate speech, profanity, social media

In May 2020, Twitter censored a post by US President Donald Trump related to the George Floyd protests with the argument that it was glorifying violence. In contrast, Facebook opted to keep the post unlabeled, which in turn generated backlash against the platform (Paul, 2020). User support and attitudes toward content moderation are key aspects that platforms should consider when designing their content moderation policies (Riedl et al., 2021). While some studies have looked at the impacts of moderation practices on users who have had their content flagged or removed (e.g. Jhaver et al., 2019a; Myers West, 2018), there is little information on how these practices affect perceptions of the large portion of users who have never engaged with moderation teams,[1] but are, nonetheless, important stakeholders for platforms.

Perceptions of the types of content that should be allowed on social media are particularly relevant when considering harmful content like hate speech and novel content moderation practices like machine learning classification. Debates over migrants, the reemergence of far-right extremism in Europe and in the United States, and online cultures of misogyny, xenophobia, and racism have highlighted the growing problem of both online hate speech (Pohjonen, 2019) and less virulent but still antagonistic content such as profanity (Chen, 2017; Krämer and Springer, 2020). In Q1 2019, Facebook removed about 4 million pieces of content due to hate speech, numbers that make it clear that human moderators are inadequate alone to handle this amount of content.

Considerable progress has been achieved in developing automated ways of detecting hateful content (e.g. Djuric et al., 2015; Zhang et al., 2018) and profanity (Gillespie, 2018). However, qualitative research (Myers West, 2018; Suzor et al., 2019) shows that even if moderation algorithms are developed, their implementation may be compromised by the way content removal decisions are communicated. For instance, a study of Wikipedia (Halfaker et al., 2012) concludes that algorithmic moderation tools are one of the likely causes for a reduction in participation on the platform, and research on Reddit shows that moderators fall short when implementing transparency and accountability principles (Juneja et al., 2020). Besides academic studies, scrutiny of algorithmic decision-making has also entered the public agenda through documentaries such as *Coded Bias* (Kantayya, 2020). We study content moderation from the perspective of a bystander,

because the technical progress made toward automated content moderation may be meaningless if the user community questions the justice and fairness of moderation. Increased perceptions of justice and fairness have been shown to reduce future norm violations on social media platforms (Tyler et al., 2021), while messages from the platforms themselves can increase these perceptions. Even if users are not involved in moderation processes directly, perceiving the moderation of other users' content as fair should still reduce deviant behavior through social learning (Bandura, 2005) and may increase bystander intervention to enforce norms (Naab et al., 2016).

Users want content moderation (da Silva, 2015), but they may also perceive it negatively (Myers West, 2018; Suzor et al., 2019). We argue that it is essential to understand better what factors play a role in user support for different types of moderation. Our study examines the interplay of three factors that have not yet been studied together: moderator type (human vs AI [artificial intelligence]), explanation for content removal (no reason for removal vs presenting general community guidelines vs giving specific removal reason), and type of content (hate speech vs profanity). We examine the influence of these factors on five outcomes related to organizational justice. For a stringent test, we conduct preregistered (https://osf.io/2n7d5/?view_only=d029d9ffa052481f83eeb9e67717a8da) online between-subjects experiments in three countries with a representative sample ($N=2870$).

The findings from this study contribute new knowledge on algorithmic aversion (Dietvorst et al., 2015) and hate speech while also offering practical recommendations for online content managers. Furthermore, the comparison of European (Netherlands and Portugal) and North American (United States) samples allows us to present findings beyond a single cultural and legal context, although generalizations are still conditioned by the fact that we are looking at a specific type of moderation, one that is managed by the platforms themselves and situated in specifically national and historical contexts.

Assessing content moderation

Despite early optimism that the Internet could be a new public sphere (Papacharissi, 2002), it soon became apparent that online spaces often devolved into wastelands of toxicity unless content was moderated (Goodman and Cherubini, 2013). While content moderation is widespread, many questions surround its implementation. Although websites publish community guidelines to clarify moderation processes, users are still often frustrated with the process and outcome of content moderation (Myers West, 2018). Furthermore, when AI is used, accountability mechanisms are less clear, and algorithms are disproportionally penalized for mistakes in user perceptions when compared to humans (Dietvorst et al., 2015).

While many studies on algorithmic aversion often frame the problem as a matter of user preferences or perceptions of accuracy (see Castelo et al., 2019), content moderation involves different facets that cannot be subsumed under a single construct. In the face of this complexity, the concept of organizational justice is an adequate starting point because it describes how people perceive the fairness of decisions that involve them (Colquitt, 2001; Lee, 2018). Moderators and users always interact in the framework of an organization, be it a social media website like Facebook or a news outlet. Organizational justice

is useful to tackle the multiple dimensions of content moderation because it encompasses distributive justice, procedural justice, and interactional justice (Colquitt et al., 2005). Rather than focusing on a single form of organizational justice, we position this work in the integrative wave of organizational justice research (Colquitt et al., 2005: 35), assessing five indicators that relate the different forms of organizational justice: outcome fairness, procedural fairness, transparency, legitimacy, and trust. These concepts capture how justice is perceived by bystanders within the organization, specifically perceptions of how others' content is moderated. While some of the concepts have been shown to be theoretically and empirically related, they, nevertheless, correspond to different dimensions of content moderation. This allows us to explore nuances that might have been missed by studies that examined only one or two constructs or a single understanding of organizational justice.

Traditionally, the idea of justice branches into procedural fairness and outcome fairness. Outcome fairness concerns perceptions regarding how just the results, consequences, and their distributions are. In contrast, a process may be perceived as fair if it includes the affected people in the process, upholds ethical and moral standards, ensures impartiality and consistency of decision-making, and provides the possibility of correcting the process (Colquitt, 2001; Ötting and Maier, 2018). While a fair process increases the likelihood that the outcome will also be perceived as fair, these are distinct concepts and dimensions.

Other concepts to consider for moderation are legitimacy and trust, because they pertain to whether people perceive the moderator as having the authority to delete content. Legitimacy is defined as the acceptance of an entity as having the right to make decisions that impact oneself and feeling obligated to follow these decisions (van Dijke et al., 2010). A legitimate moderator is perceived as having the right to make decisions about user content, and users accept the moderator's decisions (Tyler, 2006; van der Toorn et al., 2011). Trust is the conviction that an authority possesses integrity, honesty, and benevolence that render it trustworthy, so it is not perceived as exploitive (Rawlins, 2008).

Finally, in line with the fairness, accountability, transparency (FAT) framework for content moderation, we consider perceptions of transparency (Jhaver et al., 2019c). Transparency requires that an organization provide information about its actions that are "truthful, substantial and useful," so that the organization can be held accountable (Rawlins, 2008: 74). Suzor et al. (2019) emphasize the need for meaningful transparency for content moderation, such as providing sufficient explanations for why content was removed to strengthen users' trust in moderation and to ensure accountability of moderators and platforms.

Hate speech and profanity

Of aversive content posted online, hate speech has been of particular concern to platforms (Facebook, 2020), policymakers (European Commission, 2020), and scholars (Brown, 2017; Erjavec and Kovačič, 2012) because of its frequency online and negative consequences (Soral et al., 2018). Defining hate speech poses challenges because of differences in social and legal contexts.

The European Commission against Racism and Intolerance (ECRI; 2016), for example, defines hate speech as

> the use of one or more particular forms of expression—namely, the advocacy, promotion or incitement of the denigration, hatred or vilification of a person or group of persons, as well any harassment, insult, negative stereotyping, stigmatization or threat of such person or persons and any justification of all these forms of expression. (p. 16)

The definition excludes expressions that merely distress, hurt, or offend because hate speech is much more than mere dislike or bias and it tends to be discriminatory, abusive, and hostile in nature (Richardson-Self, 2018; SELMA, 2019). In contrast, US law does not outlaw hate speech, and courts have repeatedly ruled that it is constitutionally protected speech.

However, the US Constitution does not prohibit private businesses from limiting speech. Online platforms can enforce their own definitions of hate speech when deciding whether to remove content (Gagliardone, 2019). Facebook, for instance, defines it as "a direct attack on people based on what we call protected characteristics" (Facebook, 2020: n.p.).

Faced with the challenge of defining hate speech in our study in different national contexts, we focus on the two characteristics that are more consistent across contexts: (1) hate speech targets vulnerable or protected groups, and (2) hate speech implies serious and discriminative expressions against that group. It is noteworthy that the concept of protected groups exists in the United States, even though there is no legal concept of hate speech.

While scholars, platforms, and policymakers see hate speech as a particularly problematic form of online content, it is unclear whether users share this understanding and distinguish it from other forms of antagonistic content, such as profanity. Profanity is more frequent online than hate speech (Chen, 2017; Coe et al., 2014) and people readily identify it as aversive (Kenski et al., 2017; Muddiman and Stroud, 2017). Yet, profanity implies poor manners rather than threats to democracy and its basic principles (Papacharissi, 2004). Because hate speech is universally considered more harmful than profanity, it seems plausible that people would see efforts to moderate hate speech as fairer than efforts to moderate profanity. However, it is also possible that users would consider efforts to moderate profanity as fairer, because profanity is immediately noticeable as aversive (Kenski et al., 2017) while hate speech is open to interpretation based on context (Roussos and Dovidio, 2018). Due to the lack of a clear argument for directionality and the importance of context, H1 is put forward as a non-directional hypothesis:

> *H1*. There are differences in perceived procedural fairness (H1a) and outcome fairness (H1b) between the removal of hate speech and the removal of profanity content.

Algorithmic aversion

While machine-learning approaches outperform human judgment in many situations, there is still an implicit tendency for individuals to prefer human decisions or advice.

This preference has been labeled as algorithm aversion (Dietvorst et al., 2015), and it has been verified in different domains, such as medicine (Longoni et al., 2019) and employee recruitment (Diab et al., 2011), among others. Despite evidence for algorithm aversion (see Burton et al., 2020, for a review), some studies show that people prefer algorithmic judgments (Logg et al., 2019) when the focus is on decision-making and assessing outcomes. In particular, recent work on perceptions of AI focuses on fairness as the key metric to assess attitudes toward decision-makers (Helberger et al., 2020):

H2. There are differences in perceived procedural fairness (H2a) and outcome fairness (H2b) of content removal between human and algorithmic moderators.

Despite diverse findings regarding the assessment of algorithmic and human judgments, algorithms are consistently seen as underperforming when compared to humans in certain tasks. People penalize algorithms disproportionally for making mistakes (Dietvorst et al., 2015). As a relatively new mechanism in content moderation, AI moderators, therefore, may lack the legitimacy and trust that people assign to their more familiar human counterparts. Furthermore, because algorithms operate behind the scenes, their *black box* nature makes perceptions of transparency particularly important (Shin and Park, 2019).

H3. Algorithmic moderators are perceived as less trustworthy (H3a), transparent (H3b), and legitimate (H3c) than human moderators.

One key idea that underlies algorithm appreciation or aversion is task dependency, which refers to the characteristics of a given task and, in particular, the degree to which a task is seen as requiring subjective or objective reasoning (Castelo et al., 2019; Lee, 2018). While content moderation at first may appear as a single task, different types of content require moderators to employ distinct interpretative frameworks. For example, because profanity and hate speech are conceptually different, moderation decisions may draw on different abilities and resources. Because characteristics of profanity are explicit, literal, and independent of context (Kenski et al., 2017; Muddiman and Stroud, 2017), profanity can unambiguously be classified as such and filtered. In contrast, hate speech can also be implicit but does not have to be literal and, thus, cannot always be identified through specific keywords, making it highly context dependent. Deciding whether a post contains hate speech, therefore, demands a more nuanced reading and empathy for the protected group being targeted by a post. This suggests a level of subjectivity and a moral dimension (Lee, 2018) in hate speech moderation that aligns more readily with skills associated with humans, than algorithms.

These task characteristics have consequences for perceptions of the fairness of and trust in decision-makers (Lee, 2018). It can, thus, be assumed that perceptions of fairness and legitimacy as well as trust in a moderator vary with the type of content that is being scrutinized. Using this reasoning, algorithms would be preferred for objective tasks, such as recognizing and filtering profanity, and human moderators would be preferred for subjective tasks, that is, recognizing and removing hate speech. Based on these

assumptions, we hypothesize the following interaction effect, positing that the type of content influences the differences in perceptions of human and algorithmic moderators:

> *H4*. The effect of moderator type depends on the type of content being moderated, such that human moderation is perceived as more trustworthy (H4a) and legitimate (H4b) and as having more procedural fairness (H4c) and outcome fairness (H4d) when removing hate speech when compared to removing profanity.

Within the scope of algorithm aversion and appreciation, it should be considered that the ways in which AI is framed within national discourses may have an impact on how it its perceived. For instance, watching science fiction movies with AI weapon systems impacts support for autonomous weapons (Young and Carpenter, 2018). While it is clear that national differences must exist in how AI is framed within each of the countries in our study, we do not have information that would allow us to formulate expectations regarding specific country differences in perceptions. Data from the Eurobarometer (European Commission, 2019) on AI show that both the Portuguese and the Dutch are among those in Europe who agree the most that AI should be used for safety and security. Similarly, the percentage of respondents in the United States who believe that AI requires careful management is similar to the EU (European Union) average (Zhang and Dafoe, 2019). Considering that no unambiguous differences in attitudes toward AI can be found for the countries in our study, we focus our analytical strategy in testing whether effects hold across these national contexts rather than hypothesizing that effects are moderated by country.

Moderation messages

When handling hateful content online, moderation messages matter. Users often feel frustrated and confused because they do not understand or agree with the reasons why content was removed (Jhaver et al., 2019a). Even if bystanders agree with the need for content moderation (Masip et al., 2019; da Silva, 2015), lack of transparency and engagement may lead to myths and misconceptions regarding how content is handled by moderators (Suzor et al., 2019; Myers West, 2018). Users who participate in an online forum may suddenly witness a situation where content they were engaging with disappears. While the user who submitted the offending content may receive some explanation for its removal, other users are often simply confronted with a general message: "This comment was removed by a moderator because it didn't abide by our community standards" (example retrieved from *The Guardian*). The lack of specific information may in turn result in misconceptions and lack of support for moderation from these bystanders.

One answer to mitigate backfire effects of content moderation is to provide meaningful moderation messages to users. Giving additional information not only equates to more transparency, but previous studies suggest that it may increase perceptions of fairness (Jhaver et al., 2019a). As such, we expect that increasing the amount of information in a custom moderation message, indicating the specific reason why content was removed, will increase perceptions of fairness and transparency:

H5. Custom messages for content removal are perceived as having more transparency (H5a), procedural fairness (H5b), and outcome fairness (H5c), when compared to the no information condition.

In addition to the direct effect of providing information about content removal, we also expect to find interaction effects between our conditions. As mentioned above, moderating hate speech requires more contextual knowledge and is seen as more subjective (Lee, 2018). We, therefore, assume that hate speech content also increases the need for an explanation of why the content was removed:

H6. The positive effect of providing a custom message compared to no information on transparency (H6a), procedural fairness (H6b), and outcome fairness (H6c) is stronger for hate speech than for profanity.

Regarding the interaction between moderator type and explanations, previous research has found that the source of the explanation had no impact on the future behavior of users who have had their content removed (Jhaver et al., 2019c). However, different results could occur when we consider perceptions rather than behavior as a dependent variable and when no explanation is provided. Based on the argument that algorithms are *black boxes* and generally less understandable for users than another person's reasoning, we assume providing a reason as to why the content was removed will have a more positive impact for an algorithm than for a human moderator:

H7. The positive effect of providing a custom message compared to no information on transparency (H7a), procedural fairness (H7b), and outcome fairness (H7c) is stronger for the algorithmic moderator than for the human moderator.

Finally, we also test the common scenario of online platforms presenting a general moderation message that directs users to a community guidelines, policy or standards page. This approach provides more information than a simple content-removal message but only if the user is willing to visit the community guidelines. On the contrary, not indicating the specific reason for content removal opens the possibility for differing interpretations. While we expect that the general message condition will perform somewhere in the middle when compared with the remaining two explanations for removal conditions, we, nonetheless, ask the following research question:

RQ1. Are user perceptions of a general moderation message (i.e. a reference to the community guidelines) more similar to perceptions of a custom message (specific explanation for the content removal) or to perceptions of receiving no explanatory message (no further explanation for the content removal)?

Table 1 summarizes the hypothesized relationships. All hypotheses were preregistered, and decisions on directionality and dependent variable selection are grounded on the previous literature.

Table 1. Research design.

Hypothesis	Independent variable	Directionality	Dependent variable
H1	Hate speech (ref. cat. = Profanity)	+/−	Procedural fairness (H1a) Outcome fairness (H1b)
H2	Human (ref. cat. = AI)	+/−	Procedural fairness (H2a) Outcome fairness (H2b)
H3	Human (ref. cat. = AI)	+	Trust (H3a) Transparency (H3b) Legitimacy (H3c)
H4	Human (ref. cat. = AI) × Hate speech (ref. cat. = Profanity)	+	Trust (H4a) Legitimacy (H4b) Procedural fairness (H4c) Outcome fairness (H4d)
H5	Specific information (ref. cat. = No information)	+	Transparency (H5a) Procedural fairness (H5b) Outcome fairness (H5c)
H6	Specific information (ref. cat. = No information) × Hate speech (ref. cat. = Profanity)	+	Transparency (H6a) Procedural fairness (H6b) Outcome fairness (H6c)
H7	Specific information (ref. cat. = No information) × Human (ref. cat. = AI)	−	Transparency (H7a) Procedural fairness (H7b) Outcome fairness (H7c)
Control variables			
Country (ref.cat. = USA)			
Agreement with post			
Attitudes toward immigrants			

AI: artificial intelligence.

Method

This study was preregistered prior to data collection at OSF: https://osf.io/2n7d5/?view_only=d029d9ffa052481f83eeb9e67717a8da. It was approved by institutional review boards both in Europe and in the United States.

The study consists of an online between-subjects experiment targeting residents in the United States, the Netherlands, and Portugal, with data collected in April 2020. Participants were recruited from an Internet panel from Dynata. The final sample ($n = 2870$) was representative of the Internet users in each country for age, gender, education, and race/ethnicity.[2] Details on the sample for each country are provided in Supplemental Appendix 3.[3] Studying different national contexts is relevant because approaches on hate speech differ substantially in US and EU law (Bleich, 2014). In addition, even within the EU, countries are not homogeneous, underscoring the need to compare the Netherlands with Portugal to assess robustness across context. For example, in

Portugal censorship was common during the dictatorship that ended in 1974, and there are strong reactions to limiting speech, so residents there may perceive content moderation differently than in the Netherlands. In a similar way, understandings and discourses surrounding AI may vary with national contexts as well.

Stimuli, design, and procedures

The stimuli for our study consists of two parts: a social media post with a short text with offensive content (either profanity or hate speech) and a content-removal message. A total of 10 messages (five profane and five hate speech) were pretested ($n = 304$) in the three countries in our study. Two messages were selected for the final stimuli based on their perceived levels of profaneness and hatefulness, aiming at differentiating the two conditions as much as possible, that is, the hate speech message should be perceived as hateful but not profane and vice versa.[4] Messages were translated and adapted to their respective contexts. For instance, the hate speech message targeted Mexicans in the United States, Brazilians in Portugal, and foreigners in the Netherlands, to enhance the external validity of the study by referencing groups that are prominent targets of hate speech in each country. The post was edited to mimic the popular online discussion template Disqus, with the prompt referencing a "post in a social media platform." Because this template is used in diverse discussion spaces, we ensure that our results are not impacted by preconceptions that participants may have toward popular platforms, such as Facebook or Twitter. While not disclosing a specific platform limits the external validity of our findings, insofar that perceptions of moderation are always conditioned by the platform where they take place, it allows us to measure baseline effects that could be influential across contexts.

Our experiment followed a 2 (profanity vs hate speech) $\times$ 2 (algorithm vs human moderator) $\times$ 3 (explanation for removal level of detail) design. Participants first saw a hate speech or profane message and were told that the message was posted on a social media site. They were then shown a message stating that the content had been removed, along with a blurred version of the previous post to remove any ambiguity as to what content the message was referring to (see Supplemental Appendix 2). Our stimuli design aims to mimic a situation that a frequent user of an online discussion platform with a post-moderation policy might encounter, where upon revisiting a particular discussion later in time they might find that offending content has been removed. The content removal message informs the participant that the social media post was removed because it breached the platform's community guidelines, disclosing the moderator as human or algorithm. To enhance the strength of our moderator manipulation, both the human moderator (John/Joost/Pedro) and the algorithm (ModerHate) were named. We decided to present the human moderator as being a part of the platform's content moderation team, rather than a volunteer moderator, to mimic the practices of some major social media platforms like Facebook and Twitter. Although we acknowledge that some platforms, like Reddit and Discord, are primarily moderated by volunteers, we address the implications of this distinction further in the "Discussion" section. The moderation message was manipulated to disclose varying degrees of information to the user: stating that the content was removed because it violated the platform's rules (no information condition);

directing the user to the actual community guidelines, with a working link to these guidelines (general message condition); or disclosing the specific reason (profanity or hate speech) why the content was removed (custom message condition). After exposure to the message, questions related to the dependent variables followed, with control variable, demographic, and manipulation checks at the end of the survey.

Random assignment was successful, with no significant differences across conditions in terms of age, gender, race/ethnicity, and education. Manipulation checks also indicate that our manipulations of moderator type, $\chi^2(3, N=2817)=832.25$, $p<.001$, and message type $\chi^2(4, N=2739)=519.68$, $p<.001$, were successful.

Measures

Scales were adjusted to seven-point scales with answer options ranging from *completely disagree* to *completely agree*. We used confirmatory factor analysis (CFA) to validate the structure of our data. The measurement model displayed adequate fit considering the sample size: root mean square error of approximation (RMSEA) = 0.069 (90% confidence interval [CI] = [0.067, 0.071]) and comparative fit index (CFI) = 0.954 ($\chi^2=3261.79$, $df=242$, $p<.001$). After reliability checks, the measurement model was integrated in the structural model described in the "Results" section. Some of the study researchers are fluent in all three languages (English, Dutch, and Portuguese) and translated stimuli and questions. Although latent variables were used in the final model, means and standard deviations of averaged items are presented below for context.

Outcome fairness: Based on Colquitt (2001), we used four items to assess participant's perceptions of outcome fairness (e.g. "The moderator's decision is appropriate in regard to the post" and "I think it was fair to remove the post.") ($M=5.45$; $SD=1.69$; $\alpha=.96$).

Procedural fairness: Drawing from Colquitt (2001) and Koper et al. (1993), we used five statements to assess perception of the fairness of the moderation process (e.g. "The moderation process has followed ethical and moral standards.") ($M=5.43$; $SD=1.41$; $\alpha=.93$).

Legitimacy: We adapted a legitimacy scale from van der Toorn et al. (2011), using five statements to measure the legitimacy of the moderator and their content moderation practices (e.g. "The moderator has the right to evaluate posts that users make on this social networking site.") ($M=5.16$; $SD=1.29$; $\alpha=.86$).

Transparency: Transparency was assessed by adapting and using the scores of five items drawn from Rawlins (2008) to measure how transparent the participants considered the moderation process to be (e.g. "Users like me are provided with detailed information to understand the removal of the post.") ($M=5.03$; $SD=1.49$; $\alpha = .95$).

Trust: Our five-item scale for trust was adapted from Rawlins (2008) and Ohanian (1990). The items aimed to capture the extent to which participants trusted the moderator (e.g. "I think the moderator is honest when evaluating posts on this social networking site.") ($M=5.27$; $SD=1.29$; $\alpha=.92$)

Attitudes toward immigrants: Because our hate speech condition directs offensive comments at immigrants, it was important to control for this factor. To ensure a robust six-item scale to measure these attitudes, two items were drawn from Pew Research

Table 2. Model fit statistics.

	Model 1	Model 2	Model 3	Model 4
RMSEA [90% CI]	.05 [.05, .05]	.05 [.05, .05]	.05 [.05, .05]	.05 [.04, .05]
CFI	.95	.95	.95	.95
SRMR	.03	.03	.03	.02
χ^2	3453.74***	3769.01***	3839.75***	3971.45***
df	413	451	489	641
AIC	174,750.21	174,567.64	172,565.236	172,601.38

RMSEA: root mean square error of approximation; CFI: comparative fit index; SRMR: standardized root mean residual; AIC: Akaike information criterion.
Model 1: Experimental conditions and interactions between conditions.
Model 2: Experimental conditions, interactions between conditions, and country control.
Model 3: Experimental conditions, interactions between conditions, and all controls.
Model 4: Experimental conditions, interactions between conditions, all controls, and interaction between country and conditions.
***$p < .001$.

Center polling (Jones, 2019), two from the American National Election Studies, and two from Gallup polling. Three of the items are reversed, and the statements were adapted for each national context (e.g. "American identity, norms and values have been enriched thanks to the presence of immigrants.") ($M = 4.29$; $SD = 1.36$; $\alpha = .92$)

Agreement with post: Because people who agree with a post would be less likely to support its removal, we also controlled for this in our models. Agreement was measured by asking participants how much they agreed with the social media post ($M = 3.33$; $SD = 2.09$).

Model selection

To test the hypotheses, we used structural equation modeling (SEM) with the *lavaan* (Rosseel, 2012) package in R. SEM tackles the complexity of our research design while allowing us to integrate the analyses in a single research model. The latent dependent variables were included in the model as specified in the measurement model (see Supplemental Appendix 1 for details), while items on attitudes toward immigrants were averaged before their inclusion in the model.

Theoretically, we expected that control variables would improve our model. Country differences, attitudes toward immigrants, and agreement with the post should all have an impact on our dependent variables. However, we did not hypothesize interactions between our experimental conditions and controls because we did not have specific expectations that the effect of our manipulations would depend on country and attitudes. Empirically, we compared four different model specifications to assess whether adding our control variables was justified (comparing Models 2 and 3 with Model 1), and whether adding country interactions would result in an improved fit. Table 2 indicates that model fit statistics support our theoretical expectations. There are no meaningful differences in fit statistics between the four models, but the Akaike information criterion

Table 3. Structural equation model.

	Outcome fairness	Procedural fairness	Legitimacy	Transparency	Trust
Hate speech	0.92 (0.11)***	0.59 (0.10)***	0.45 (0.10)***	0.37 (0.11)**	0.29 (0.09)**
Human	−0.02 (.11)	−0.06 (.10)	0.02 (.10)	−0.35 (.11)**	−0.07 (.09)
General message	−0.28 (0.12)*	−0.17 (0.10)	−0.16 (0.10)	0.04 (0.11)	−0.26 (0.10)**
Custom message	−0.05 (0.12)	0.03 (0.10)	0.01 (0.10)	0.41 (0.11)***	−0.08 (0.10)
Hate speech × human	−0.05 (0.11)	−0.02 (0.10)	−0.05 (0.10)	0.18 (0.11)	0.04 (0.09)
General message × human	0.29 (0.14)*	0.20 (0.12)	0.18 (0.12)	0.28 (0.13)*	0.19 (0.11)
Custom message × human	0.22 (.14)	0.19 (.12)	0.13 (.12)	0.33 (.13)*	0.24 (.11)*
Hate speech × general message	−0.00 (0.14)	−0.01 (0.12)	0.04 (0.12)	0.12 (0.13)	0.18 (0.11)
Hate speech × custom message	−0.06 (0.14)	−0.11 (0.12)	−0.00 (0.12)	−0.09 (0.13)	0.05 (0.11)
Agreement with post	−0.13 (0.01)***	−0.10 (0.01)***	−0.08 (0.01)***	−0.04 (0.01)**	−0.06 (0.01)***
Attitudes immigrants	0.18 (0.02)***	0.16 (0.02)***	0.13 (0.02)***	0.12 (0.02)***	0.13 (0.02)***
Netherlands	0.34 (0.07)***	0.30 (0.06)***	0.33 (0.06)***	–0.12 (0.07)	0.28 (0.06)***
Portugal	0.67 (0.07)***	0.61 (0.06)***	0.41 (0.06)***	0.20 (0.07)**	0.52 (0.06)***

AI: artificial intelligence.
Values in each cell are presented as unstandardized estimate (standard error).
Reference category for hate speech is profanity, reference category for human is AI, reference category for general and custom message is no information, and reference category for country is the United States.
***$p < .001$, **$p < .01$, *$p < .05$.

(AIC) suggests that Model 3, having the experimental conditions, interactions between conditions and controls as predictors for our five dependent variables, is superior. Fit statistics indicated an adequate fit of the model to our data: RMSEA = 0.051 (90% CI = [0.050, 0.053]) and CFI = 0.950 ($\chi^2 = 3839.75$, $df = 489$, $p < .001$), even though the chi-square test is significant due to our large sample size.[5]

Results

While SEM models are usually represented visually, the size and complexity of our model make it easier to display the results in a tabular form (Table 3). In support of H1, which predicted differences in perceived procedural fairness (1) and outcome fairness (2) between removal of hate speech and profanity content, we see that the type of content being moderated has an impact on both dependent variables. In fact, outcome fairness (US $\Delta M = 1.24$; NL $\Delta M = 1.13$; PT $\Delta M = 0.80$, all $p < .001$) and procedural fairness (US $\Delta M = 0.84$; NL $\Delta M = 0.71$; PT $\Delta M = 0.49$, all $p < .001$) are consistently higher on the hate speech condition for all three countries.

In contrast, there were no significant differences in outcome or procedural fairness between the two forms of human and algorithmic moderation tested in our study, thus rejecting H2. Addressing H3, which predicted that algorithmic moderators would be perceived as (a) less trustworthy, (b) transparent, and (c) legitimate than human moderators, we found no significant effects of moderator type on legitimacy and trust. However, for H3b, we found the opposite effect of what we predicted: The algorithmic moderator was

rated as more transparent than the human moderator ($B=-.35$, $p=.001$), not less. H3 is rejected.

H4 predicted an interaction between content type and moderator type, such that humans are seen as more just when removing hate speech. However, as shown in Table 3, none of the interaction terms involving content type was significant. This means that although there was a direct effect of content type on our dependent variables, this variable did not condition participants' preference for a human or algorithmic moderator, at least in the form of moderation tested in our experiment. Thus, H4 is rejected.

H5 stated that custom messages for content removal have higher procedural fairness (H5a), outcome fairness (H5b), and transparency (H5c), when compared to the no information condition. It is also related to content removal explanations. RQ1 asked whether user perceptions of general moderation message were closer to a custom message or to receiving no information in the message. Regarding H5, a custom message, indicating the specific reason why content was removed, only performed better than a no information message in terms of transparency. Therefore, the data support H5c, but H5a and H5b are rejected. Regarding RQ1, it is interesting to note that a general message, represented by the common practice of directing the users to community guidelines, is outperformed in terms of outcome fairness and trust by the no information condition. When probing the significant interaction term for outcome fairness between the general message and the moderator type conditions, we see that the low performance of the general message condition happens mostly in the AI condition.

H6 posited that that there would be an interaction between message type and content type, such that an explanation would have a stronger positive impact on fairness when removing hate speech when compared to profanity. As mentioned, none of the interaction terms containing the content type condition were significant, meaning that H6 is rejected. In turn, H7 hypothesized a significant interaction effect for the moderator type and message type conditions, such that algorithms would benefit more from the effect of providing a custom message. Table 3 shows a significant interaction effect for transparency in this regard, illustrated through the estimated marginal means in Figure 1. However, as shown by the estimated marginal means, providing a custom message benefits the transparency of a human moderator more than an algorithmic moderator, particularly when compared to the no information condition. This goes against our initial expectation of AI benefiting more from additional information, and it means that H7 is rejected.

Finally, even though we tested a specific form of moderation, it should be noted that the conclusions drawn are robust across different countries. Not only did Model 4 (with country interactions) have worse fit statistics, but also most interaction effects between experimental conditions and country were nonsignificant. The only exception was the interaction between Portugal and the hate speech condition, which had a significant negative effect on outcome fairness ($B=-.36$, $p=.011$), procedural fairness ($B=-.29$, $p=.015$), and legitimacy ($B=-.36$, $p=.002$). This is in line with the results from our pretest, which indicated that the distinction between profanity and hate speech is less clear for the Portuguese.

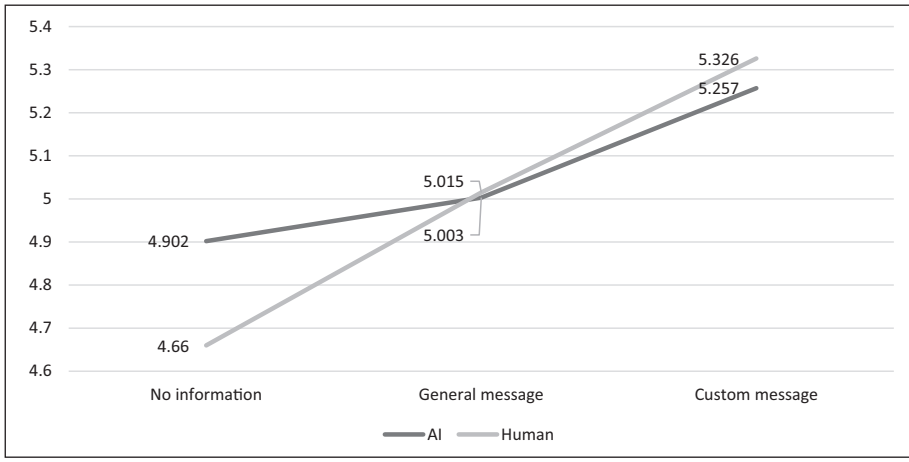


Figure 1. Estimated marginal means for transparency.
Values for attitude toward immigrants held constant at 4.29 and values for agreement with post held constant at 3.33.

Discussion

We explored how different factors affect perceptions of online content moderation and whether these effects were robust across countries. By doing so, our findings contribute to theories on algorithmic aversion (Dietvorst et al., 2015) and hate speech processing (Kenski et al., 2017), and provide practical insights into online platforms for handling hate speech and profanity.

Concerning algorithmic aversion, our study shows that there are significant differences in terms of how individuals perceive the forms of human and AI moderation tested by our experiment, but these differences are far from linear. While previous studies suggested that the type of task influences how people perceive algorithms (Castelo et al., 2019), our study concludes that the type of content being deleted (profanity vs hate speech) has no effect on perceptions of AI moderators. Our findings suggest that it is likely that bystanders to content moderation see it as a single task, and do not believe that humans or algorithms are more suitable to moderate specific types of content in settings similar to our experiment. While our stimuli exposed participants to a situation that is commonly encountered online, it should be noted that differences between moderator types may emerge in contexts where the outcome is uncertain or unknown, unlike our setting where a decision was presented to participants. Studies on algorithmic aversion have shown that the actual performance of an algorithm is key to perceptions (Castelo et al., 2019; Dietvorst et al., 2015). Our study presented participants with a situation where an algorithm correctly identified the presence of hate speech and profanity in a text, which may have reinforced their beliefs of the capabilities of AI. While this is true for all conditions, if the outcome were not known, it would be possible that participants would show a preference toward one of the moderator types, since algorithms tend to be penalized more heavily by mistakes. Likewise, it should be noted that, while these

findings apply to bystanders, one should be cautious in generalizing these findings to individuals who were personally subjected to moderation. Specifically, due to the third-person effect (Davison, 1983), bystanders may be more inclined to act against objectionable content than the publishers of the said content, a hypothesis that is supported by findings regarding Internet pornography (Sun et al., 2008). This is especially relevant considering studies that show how the third-person effect also applies to hate speech perceptions (Guo and Johnson, 2020).

Although content type was not a significant moderator in this context, our findings highlight how algorithmic perceptions depend on how a decision is communicated and on the dependent variable being measured. For transparency, our study found evidence for algorithmic appreciation, especially when no additional information on the decision is given to the participant. While algorithms are generally assumed to be *black boxes*, no previous algorithm aversion studies that we are aware of have transparency as a dependent variable to verify this (Castelo et al., 2019). The fact that this difference only occurred for the no information condition hints that, when provided with additional information on why the decision was made, information supersedes moderator type and erodes differences in perceived transparency. This finding aligns with the results from Jhaver et al. (2019c), who found no differences in posting behavior between users who received content moderation explanations by algorithms or volunteer humans.

One important caveat regarding our findings on the role of human moderators is that our study operationalized human moderation being enforced by a platform's employee and not by volunteer moderators, who are the focus of other studies on content moderation such as the one referenced above (Jhaver et al., 2019c). Volunteer moderators may have more autonomy regarding content deletion decisions, and research shows that their views on transparency are not uniform (Juneja et al., 2020), which means that perceptions regarding this type of moderation may also be distinct from what we encountered. In addition, volunteer moderators may have a relationship with community members that hired moderators lack, influencing dimensions such as trust and legitimacy from those community members. Regarding our operationalization of moderation, it should also be noted that while we chose to clearly distinguish between human and AI moderation, actual practices, like Reddit's Automoderator (Jhaver et al., 2019b), may embody a collaboration between human volunteers and AI tools, effectively introducing a hybrid type of content moderation.

The importance of accounting for different dependent variables and information conditions when assessing algorithm aversion is strengthened by our finding that AI has lower outcome fairness than humans, when users are directed toward community guidelines. This difference is not unexpected because previous research indicates that algorithmic aversion is partly due to the additional cognitive load required to follow algorithmic advice when compared to humans (Burton et al., 2020). However, most previous studies had different processes associated with following human or algorithmic advice, while our study kept the actual work required from the user consistent (visiting the community guidelines). A plausible explanation for our finding, therefore, lies in expectancy violation theory (Burgoon, 1993; Kizilcec, 2016). One key advantage that algorithms seem to have related to human agents lies in their perceived personalization (Eslami et al., 2018). However, sending a user to general guidelines disrupts that

expectation of personalization, therefore, erasing the advantage that algorithms would have in terms of outcome fairness.

Our study not only highlights how perceptions on moderators and moderation depend on how decisions are communicated, but also that the direction of the effects is contingent on the dependent variable that is measured. The opposite effects found in our study illustrate why different scholars found evidence for both algorithm aversion (Dietvorst et al., 2015) and appreciation (Logg et al., 2019), and why the discussion on perceptions of AI is more complex than one agent simply outperforming the other.

Concerning the differences between deleting online hate speech and profanity, our results show that deleting the former is clearly seen as more just. Like most findings in our study, this difference is consistent across all three countries, despite distinct legal frameworks. This is an encouraging finding because scholars (Papacharissi, 2004) tend to see hate speech as more harmful for democracy than profane language. While hate speech may be in the eye of the beholder (Roussos and Dovidio, 2018), users do make a distinction between this type of content and other forms of harmful speech when assessing what is permissible online.

While specific to the form of content moderation that was tested, our findings mean that transitioning to AI-reliant content moderation systems does not necessarily come at a cost to user perceptions, at least for bystanders to moderation. In fact, AI moderation may be beneficial in terms of transparency when it is not feasible to provide users with additional information about why content is removed. However, practitioners using AI should be wary of sending users to a community guidelines page because this had detrimental consequences for how bystanders perceived the fairness of moderation. Due to the severity of hate speech, any change in these outcomes, even if small and with limited external validity, may be consequential. We believe community managers, platforms, and policymakers should carefully consider any contextual factors that make the presence of hate speech online more acceptable. Finally, our results should be considered in light of the social and historical contexts of each platform, since these can have a significant impact on how users perceive moderation (Schoenebeck et al., 2020). While we chose not to disclose any specific platform in our experiment, practitioners should also reflect on the perceptions of their users and how previous interactions may condition their engagement with moderation systems.

Limitations and future research

We examine participant perceptions regarding content removal from the perspective of a bystander; therefore, while it is tempting to generalize these findings to those who personally have had their hateful or profane content removed, effects may be stronger, and even distinct in directionality for these cases. However, operationalizing first-person effects in a controlled between-subjects experimental setting with random assignment, while possible, would have been particularly challenging due to external validity (i.e. replicating the natural circumstances where participants would post hateful messages) and ethical concerns (i.e. prompting participants to produce hate speech messages). Nevertheless, we are encouraged by the fact that some of our findings, such as the lack of difference in perceived transparency when explanations are given by humans and AI,

are aligned with results that looked at behavior on actual platforms like Reddit (Jhaver et al., 2019c). Furthermore, while online survey experiments on perceptions of the removal of one's own content may be difficult to operationalize, online field experiments testing similar factors and perceptions are possible and yield relevant results (Marias and Mou, 2018). In our case, the bystander perspective allows us to contribute to the literature on algorithmic aversion, by looking at perceptions of algorithms performing a task that has been increasingly delegated to them. In addition, the success of content moderation systems in curtailing the spread of online hate speech also depends on a general support for these systems, which cannot be limited to users who had firsthand experience with content deletion. Given the severity of the hateful content in our experiment, any shifts in support for moderation in this instance should be considered carefully.

In addition, it is important to note that the results of this study, despite its cross-national component, are time and context bound. As technical developments and popular culture portrayals of AI develop, the way individuals perceive and relate to machine learning is likely to change, including content management and moderation. However, while work on algorithmic bias (Raji and Buolamwini, 2019) and documentaries such as *Coded Bias* may eventually lead to changes in perceptions, this article provides not only a relevant baseline for future longitudinal studies, but also contributes to unveiling some of the theoretical mechanisms that may influence how perceptions and engagement with AI content moderation develops.

Our findings regarding multiple dependent variables and the wording of moderation explanations can be used by other scholars to further explore the nuances in algorithmic aversion and appreciation in specific contexts and platforms. A research design that incorporates mediation and causality, for instance, would be able to clarify the role of transparency and trust have in perceptions of decision-making outcomes. Overall, even though our study only tested a limited set of the many forms that content moderation can take, our findings suggest that future research in this strand should consider the implications of dependent variable selection and wording of the stimuli.

Declaration of conflicting interest

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: This study was made possible through a Facebook research award. Since the topic of the study relates directly to Facebook's activity, there could be a potential conflict of interest. However, Facebook was not involved at any stage of the design, execution, or reporting of the study. The funds were awarded as an unrestricted gift, and all research was carried out independently by the authors without any external interference or influence.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors received the financial support from Facebook for this project.

ORCID iDs

João Gonçalves https://orcid.org/0000-0002-8948-0455
Ina Weber https://orcid.org/0000-0002-3501-6905
Gina M Masullo https://orcid.org/0000-0002-4909-2116
Joep Hofhuis https://orcid.org/0000-0001-7531-8644

Supplemental material

Supplemental material for this article is available online.

Notes

1. For instance, according to the Facebook Community Standards Enforcement Report for Q4, less than 0.1% of posts are removed because of hate speech.
2. Race/ethnicity questions were not asked to the Portuguese since these are forbidden by the Portuguese Constitution. In the Netherlands, the country of birth of the mother/father was asked as proxy for race and ethnicity, since US categories do not translate well to this context.
3. A total of 4603 participants filled in the survey. Participants who did not answer ($n = 615$) or who failed the attention check ($n = 878$), completed the survey in less than 48% of the median duration ($n = 213$), and straight-liners ($n = 27$) were excluded from the analysis. This resulted in a sample of 2870 participants, ($n = 902$ in the United States; $n = 975$ in the Netherlands; and $n = 993$ in Portugal).
4. Respondents were asked the extent to which each of 10 pretested messages contained profanity or hate speech. The message selected for the profanity condition was rated as significantly ($p < .001$) more profane ($M = 3.95$) than the message for the hate speech condition ($M = 3.12$). Similarly, participants reported a significantly ($p < .001$) higher presence of hate speech for the message of the hate speech condition ($M = 4.35$) than for the message on the profanity condition ($M = 3.66$).
5. The only significant change in estimates between Model 3 and a model accounting for interactions between the conditions and the control variables is that the negative effect of a general message on procedural fairness is significant with $p = .022$. All other significance levels ($\alpha = .050$) and effect directions for the experimental conditions and their interactions are robust across models.

References

Bandura A (2005) The evolution of social cognitive theory. In: Smith KG and Hitt MA (eds) *Great Minds in Management*. Oxford: Oxford University Press, pp. 9–35.

Bleich E (2014) Freedom of expression versus racist hate speech: explaining differences between high court regulations in the USA and Europe. *Journal of Ethnic and Migration Studies* 40(2): 283–300.

Brown A (2017) What is so special about online (as compared to offline) hate speech? *Ethnicities* 18(3): 297–326.

Burgoon JK (1993) Interpersonal expectations, expectancy violations, and emotional communication. *Journal of Language and Social Psychology* 12(1–2): 30–48.

Burton JW, Stein MK and Jensen TB (2020) A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making* 33(2): 220–239.

Castelo N, Bos MW and Lehmann DR (2019) Task-dependent algorithm aversion. *Journal of Marketing Research* 56(5): 809–825.

Chen GM (2017) *Online Incivility and Public Debate: Nasty Talk*. New York: Palgrave Macmillan.

Coe K, Kenski K and Rains SA (2014) Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication* 64: 658–679.

Colquitt JA (2001) On the dimensionality of organizational justice: A construct validation of a measure. *Journal of Applied Psychology* 86(3): 386–400.

Colquitt JA, Greenberg J and Zapata-Phelan CP (2005) What is organizational justice? A historical overview. In: Greenberg J and Colquitt JA (eds) *Handbook of Organizational Justice*. New York: Lawrence Erlbaum Associates, pp. 3–56.

da Silva MT (2015) What do users have to say about online news comments? Readers' accounts and expectations of public debate and online moderation: A case study. *Participations: Journal of Audience and Reception Studies* 12: 32–44.

Davison WP (1983) The third-person effect in communication. *Public Opinion Quarterly* 47(1): 1–15.

Diab DL, Pui SY, Yankelevich M, et al. (2011) Lay perceptions of selection decision aids in U.S. and non-U.S. samples. *International Journal of Selection and Assessment* 19(2): 209–216.

Dietvorst BJ, Simmons JP and Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144: 114–126.

Djuric N, Zhou J, Morris R, et al. (2015) Hate speech detection with comment embeddings. In: *Proceedings of the 24th international conference on World Wide Web*, Florence, 18–22 May 2015, pp. 29–30. New York: ACM.

ECRI (2016) ECRI general policy recommendation no.15 on combating hate speech. *Council of Europe*. Available at: https://rm.coe.int/ecri-general-policy-recommendation-no-15-on-combating-hate-speech/16808b5b01 (accessed June 2021).

Erjavec K and Kovačič MP (2012) "You don't understand, this is a new war!" Analysis of hate speech in news web sites' comments. *Mass Communication and Society* 15(6): 899–920.

Eslami M, Krishna Kumaran SR, Sandvig C, et al. (2018) Communicating algorithmic process in online behavioral advertising. In: *Proceedings of the 2018 CHI conference on human factors in computing systems*, Montreal, QC, Canada, 21–26 April 2018, pp. 1–13. New York: ACM.

European Commission (2019) Standard Eurobarometer 92: Europeans and artificial intelligence. Available at: http://ec.europa.eu/commfrontoffice/publicopinion (accessed June 2021)

European Commission (2020) The EU Code of conduct on countering illegal hate speech online. Available at: https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en (accessed June 2021)

Facebook (2020) Hate speech. Available at: https://www.facebook.com/communitystandards/hate_speech (accessed June 2021).

Gagliardone I (2019) Defining online hate and its "public lives": What is the place for "extreme speech"? *International Journal of Communication* 13: 3068–3087.

Gillespie T (2018) *Custodians of the Internet: Platforms, Content Moderation and the Hidden Decisions That Shape Social Media*. New Haven, CT: Yale University Press.

Goodman E and Cherubini F (2013) *Online Comment Moderation: Emerging Best Practices*. Frankfurt: World Association of Newspapers and News Publishers.

Guo L and Johnson BG (2020) Third-person effect and hate speech censorship on Facebook. *Social Media + Society* 6(2): 1–12.

Halfaker A, Geiger RS, Morgan JT, et al. (2012) The rise and decline of an open collaboration system: how Wikipedia's reaction to popularity Is causing its decline. *American Behavioral Scientist* 57(5): 664–688.

Helberger N, Araujo T and de Vreese CH (2020) Who is the fairest of them all? Public attitudes and expectations regarding automated decision-making. *Computer Law & Security Review* 39: 105456.

Jhaver S, Appling DS, Gilbert E, et al. (2019a) "Did you suspect the post would be removed?" Understanding user reactions to content removals on Reddit. In: *Proceedings of the ACM on human-computer interaction*, Glasgow, 4–7 May 2019, pp. 1–33. New York: ACM.

Jhaver S, Birman I, Gilbert E, et al. (2019b) Human-machine collaboration for content regulation: The case of Reddit automoderator. *ACM Trans. Computer-human Interactions* 26(5): 1–35.

Jhaver S, Bruckman A and Gilbert E (2019c) Does transparency in moderation really matter?: User behavior after content removal explanations on reddit. In: *Proceedings of the ACM on Human-computer Interaction*, Glasgow, 4–7 May 2019, pp. 1–27. New York: ACM.

Jones B (2019) Majority of Americans continue to say immigrants strengthen the U.S. *Pew Reseach, Pew Research*. Available at: https://www.pewresearch.org/fact-tank/2019/01/31/majority-of-americans-continue-to-say-immigrants-strengthen-the-u-s/ (accessed June 2021).

Juneja P, Subramanian DR and Mitra T (2020) Through the looking glass: Study of transparency in Reddit's moderation practices. In: *Proceedings of the ACM Human-computer Interactions*, Honolulu, HI (Cancelled Due to COVID)19, pp. 1–35. New York: Association for Computing Machinery.

Kantayya S (2020) *Coded Bias*. Netflix.

Kenski K, Coe K and Rains SA (2017) Perceptions of uncivil discourse online: an examination of types and predictors. *Communication Research* 47(6): 795–814. https://doi.org/10.1177/0093650217699933

Kizilcec RF (2016) How much information? Effects of transparency on trust in an algorithmic interface. In: *Proceedings of the 2016 CHI conference on human factors in computing systems*, San Jose, CA, 7–12 May 2016, pp. 2390–2395. New York: ACM.

Koper G, Van Knippenberg D, Bouhuijs F, et al. (1993) Procedural fairness and self-esteem. *European Journal of Social Psychology* 23(3): 313–325. https://doi.org/10.1002/ejsp.2420230307

Krämer B and Springer N (2020) Ontology of opposition online: Representing antagonistic structures on the Internet. *Studies in Communication and Media* 9(1): 35–61.

Lee MK (2018) Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5(1): 1–16.

Logg JM, Minson JA and Moore DA (2019) Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151: 90–103.

Longoni C, Bonezzi A and Morewedge CK (2019) Resistance to medical artificial intelligence. *Journal of Consumer Research* 46(4): 629–650.

Masip P, Ruiz-Caballero C and Suau J (2019) Active audiences and social discussion on the digital public sphere. Review article. *El Profesional De La Información* 28(2): 1–40.

Marias JN and Mou M (2018) Civil Servant: community-led experiments in platform governance. In: *Proceedings of the 2018 CHI conference on human factors in computing systems*, Montreal, Qanada, 21–26 April 2018, pp. 1–13. New York, NY: Association for Computing Machinery. https://doi.org/10./3173574.3173583.

Muddiman A and Stroud NJ (2017) News values, cognitive biases, and partisan Incivility in comment sections. *Journal of Communication* 67: 586–609.

Myers West S (2018) Censored, suspended, shadowbanned: user interpretations of content moderation on social media platforms. *New Media & Society* 20(11): 4366–4383.

Naab TK, Kalch A and Meitz TGK (2016) Flagging uncivil user comments: effects of intervention information, type of victim, and response comments on bystander behavior. *New Media & Society* 20(2): 777–795.

Ötting SK and Maier GW (2018) The importance of procedural justice in human–machine interactions: intelligent systems as new decision agents in organizations. *Computers in Human Behavior* 89: 27–39.

Ohanian R (1990) Construction and validation of a scale to measure celebrity endorsers' perceived expertise, trustworthiness, and attraction. *Journal of Advertising* 19(3): 39–52.

Papacharissi Z (2002) The virtual sphere: the internet as a public sphere. *New Media & Society* 4: 9–27.

Papacharissi Z (2004) Democracy online: civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society* 6: 259–283.

Paul K (2020) Zuckerberg: Facebook will review policies after backlash over Trump posts. *The Guardian*, 6 June. Available at: https://www.theguardian.com/technology/2020/jun/05/mark-zuckerberg-facebook-trump-policies-review

Pohjonen M (2019) A comparative approach to social media extreme speech: online hate speech as media commentary. *International Journal of Communication* 13: 3088–3103.

Raji I and Buolamwini J (2019) Actionable auditing: investigating the impact of publicly naming biased performance results of commercial AI products. In: *conference on artificial intelligence, ethics, and society*, Honolulu, HI, 27–28 January 2019, Palo Alto, CA: Association for the Advancement of Artificial Intelligence. https://hdl.handle.net/1721.1/123456

Rawlins B (2008) Measuring the relationship between organizational transparency and employee trust. *Public Relations Journal* 2(2): 1–21.

Richardson-Self L (2018) Woman-hating: on misogyny, sexism, and hate speech. *Hypatia* 33(2): 256–272.

Riedl MJ, Whipple KN and Wallace R (2021) Antecedents of support for social media content moderation and platform regulation: the role of presumed effects on self and others. *Information, Communication & Society*. Epub ahead of print 26 January 2021. DOI: 10.1080/1369118X.2021.1874040

Rosseel Y (2012) lavaan: an R package for structural equation modeling. *Journal of Statistical Software* 48(2): 1–36. http://www.jstatsoft.org/v48/i02/

Roussos G and Dovidio JF (2018) Hate speech is in the eye of the beholder: the influence of racial attitudes and freedom of speech beliefs on perceptions of racially motivated threats of violence. *Social Psychological and Personality Science* 9(2): 176–185.

Schoenebeck S, Haimson OL and Nakamura L (2020) Drawing from justice theories to support targets of online harassment. *new media & society* 23(5): 1278–1300.

SELMA (2019) *Hacking Online Hate: Building an Evidence Base for Educators*. Available at: www.hackinghate.eu (accessed June 2021)

Shin D and Park YJ (2019) Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior* 98: 277–284.

Soral W, Bilewicz M and Winiewski M (2018) Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior* 44(2): 136–146.

Sun Y, Shen L and Pan Z (2008) On the behavioral component of the third-person effect. *Communication Research* 35(2): 257–278.

Suzor NP, West SM, Quodling A, et al. (2019) What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication* 13: 1526–1543.

Tyler T, Katsaros M, Meares T, et al. (2021) Social media governance: can social media companies motivate voluntary rule following behavior among their users? *Journal of Experimental Criminology* 17(1): 109–127.

Tyler TR (2006) Psychological perspectives on legitimacy and legitimation. *Annual Review of Psychology* 57(1): 375–400.

van der Toorn J, Tyler TR and Jost JT (2011) More than fair: outcome dependence, system justification, and the perceived legitimacy of authority figures. *Journal of Experimental Social Psychology* 47(1): 127–138.

van Dijke M, De Cremer D and Mayer DM (2010) The role of authority power in explaining procedural fairness effects. *Journal of Applied Psychology* 95(3): 488–502.

Young KL and Carpenter C (2018) Does science fiction affect political fact? Yes and no: a survey experiment on "killer robots.." *International Studies Quarterly* 62(3): 562–576.

Zhang B and Dafoe A (2019) *Artificial Intelligence: American Attitudes and Trends*. Center for the Governance of AI, Future of Humanity Institute, University of Oxford. Available at: https://governanceai.github.io/US-Public-Opinion-Report-Jan-2019/index.html (accessed June 2021).

Zhang Z, Robinson D and Tepper J (2018) Detecting hate speech on Twitter using a convolution-GRU based deep neural network. In: *European semantic web conference*, Crete, 3–7 June 2018. Cham: Springer.

Author biographies

João Gonçalves is an assistant professor (tenured) at the Department of Media & Communication of Erasmus University Rotterdam. He obtained his PhD in Communication Studies from the University of Minho (Portugal), and his research focuses on artificial intelligence, audience engagement, and political communication.

Ina Weber is a doctoral student at the Department of Communication Studies at the University of Antwerp. Her research focuses on the risk factors that facilitate hate speech and how to use technology to mitigate these risks. She obtained her MA degree in Media, Culture, and Society from Erasmus University Rotterdam.

Gina M. Masullo is an associate professor in the School of Journalism and Media and Associate Director of the Center for Media Engagement in the Moody College of Communication at The University of Texas at Austin. Her research focuses on how the digital space both connects and divides people and how that influences society, individuals, and journalism.

Marisa Torres da Silva is an assistant professor (tenured) at the School of Social Sciences and Humanities, Universidade Nova de Lisboa (NOVA FCSH). She obtained her PhD in Communication Studies at the same institution and her research focuses on the relationship between journalism, democracy and audiences, online hate speech, news and civic literacy, news consumption and audience research, gender and journalism, and cultural journalism.

Joep Hofhuis is an assistant professor at ESHCC's Department of Media and Communication, where he specializes in intercultural and organizational communication. He was awarded a PhD in social and organizational psychology (University of Groningen, 2012), based on his research on cultural diversity in the workplace.