"I will defend your right to free speech, provided I agree with you": How social media users react (or not) to online out-group aggression

Abstract

Social networking sites (SNS) routinely ban aggressive users. Such bans are sometimes perceived as a limitation to the right to free speech. While research has examined SNS users' perceptions of online aggression, little is known about how observers make trade-offs between free speech and the desire to punish aggression. By focusing on reactions to a SNS ban, this study explores under what circumstances users consider the protection of the right to free speech as more important that the suppression of aggression. We propose a model of moderated mediation that explains under what circumstances online aggression increases the acceptance of a ban. When posts display aggression, the ban is less likely to be perceived as violating free speech and as unfair. Consequently, aggression reduces the likelihood that users will protest through negative word of mouth. Moreover, users protest against a social networking site ban only when this affects an in-group user (rather than an out-group user). This in-group bias, however, diminishes when an in-group aggressor targets a high warmth out-group user. The study raises managerial implications for the effective management of aggressive interactions on SNS and for the persuasive communication of a decision to ban a user engaging in aggressive behavior.

Keywords: free speech; online aggression; anger; social networking site bans; unfairness; negative word of mouth.

'It is better to debate a question without settling it, than to settle a question without debating it." Joseph Joubert

Introduction

On November 23, 2018, the Canadian journalist Meghan Murphy was banned permanently from Twitter. She had posted several comments on how society should view transgenderism. Murphy posted that "men aren't women" and asked: "What is the difference between a man and a trans woman?" She also used a male pronoun to refer to a transgender who identifies as a woman. Twitter considered such statements hateful speech, because they degrade someone based on their gender identity (BBC News, 2019). Murphy protested and launched a lawsuit for what she sees as a dangerous violation of the right to free speech (Wells, 2019).

The case illustrates the important role that social networking sites (SNS) play in the promotion (or hindrance) of the right to free speech. The growing importance of SNS as forums for debate of social and political issues increases societal scrutiny on how they handle the controversies that such debates sometimes generate (Klein, 2018; Malik, 2018). Indeed, SNS have often willingly assumed this mantle, as in the case of Twitter, which boasted of being "the free speech wing of the free speech party" (Halliday, 2012). However, sometimes SNS ban users who have violated their policies. In such circumstances, SNS might communicate their banning decision in order to remind other users of their rules of conduct and to justify their choices (Walawalkar, 2020). In other cases, SNS users might become aware of the ban and react negatively when such decisions draw the public's attention (Klein, 2018). The fact that more than 70% of Americans, and more than 80% of Republican-leaning voters, believe that SNS intentionally censure opinions they do not agree with (Smith, 2018) testifies to users' skepticism toward SNS bans. At the same time, users wish to support the victims of online aggression (Hershcovis & Bhatnagar, 2017). This results into a managerial dilemma for SNS: how to balance users'

interest in exercising the right to free speech with the need to punish aggressive behavior (Antoci et al., 2019).

Past research has focused on observers' perceptions of aggression (Anderson et al., 2018; Barnidge, 2017; Hershcovis & Bhatnagar, 2017). Online aggression is intentional and targets one specific person (i.e. the victim) (Hershcovis & Bhatnagar, 2017). Yet research to date has overlooked how users react to information about a ban that seeks to halt aggressive behavior. Users generally have negative emotional reactions to online aggression (Bacile, Wolter, Allen, & Xu, 2018) and dislike aggressive displays on SNS (Anderson et al., 2018; Barnidge, 2017). Consequently, bans aimed at reducing aggression should be unanimously appreciated. The issue is however more complex if we consider users' concerns about preserving their right to free speech. From this perspective, users might react negatively to a ban perceived as a violation to free speech (Etzioni, 2019). Extant research however has not considered users' reactions to a ban intended to limit aggression on SNS. By examining directly users' reactions to SNS bans, this study focuses on how observers make trade-offs between their willingness to protect free speech and the desire to suppress aggression. It develops our understanding of the circumstances under which users will consider the right to free speech as more (or less) important than the suppression of online aggressive behavior. Specifically, we study under what conditions observers are more (or less) likely to oppose a ban from a social networking site.

Addressing this research gap is pivotal for SNS because scholars in law and ethics disagree on what is offensive or harmful speech, and on whether such forms of speech should be restricted (DuMont, 2016; Etzioni, 2019; Nielsen, 2018). Concurrently, there is growing recognition that high levels of online incivility are detrimental to the well-being of users (Bacile, Wolter, Allen, & Xu, 2018; Muddiman, 2017) and have negative aggregate consequences on social and political debates (Antoci et al., 2019; Gervais, 2019; Soral, Bilewicz, & Winiewski, 2017). Empirical evidence on how users evaluate online aggression

and potential reductions to free speech is crucial for informing the decisions of SNS and policy makers concerning the introduction of codes of conduct which are effective at reducing aggression while meeting the approval of SNS users.

Building on deontic justice theory, we develop and test a model of moderated mediation that explains users' reactions to a social networking site ban. We predict and demonstrate that perceptions that a ban violates free speech and is unfair lead to negative word of mouth in protest against the ban. However, observers are less likely to perceive the ban as unfair or as a violation to free speech when the banned user posts aggressive content. Moreover, observers of aggression are much less likely to protest when the ban concerns a user belonging to an out-group. This protective in-group bias is indicative of a desire to protect in-group members and in-group opinions from the SNS ban. Such bias is reduced when the in-group user has attacked an out-group member perceived as high in warmth. Observers are less likely to protest a ban of an in-group aggressor targeting a warm out-group user.

To the best of our knowledge, this is the first study to model the factors that influence the acceptance of restrictions on free speech by SNS. We contribute to the literature on the positions taken by consumers in relation to the protection of the right to free speech (Etzioni, 2019; Klein, 2018; Nielsen, 2018). In particular, we demonstrate that the ban's perceived violation of free speech and perceived unfairness lead observers to engage in negative word of mouth. Our evidence shows that, for SNS users, restrictions to free speech represent a form of controversial or ethically questionable behavior that deserves punishment (Antonetti & Anesa, 2017; Carillat, O'Rourke, & Plourde, 2019; Xie & Bagozzi, 2019).

The study also contributes to research on the social perception of aggression (Antoci et al., 2019; Hershcovis, 2011; Muddiman, 2017) in three ways. First, consistent with deontic justice theory (Folger, 2001), we show that aggression reduces opposition to user bans. This implies that the decision of a social networking site to ban users gains legitimacy among observers of

aggression. This is the first investigation to show, through the application of deontic justice theory (Folger, 2001) to SNS' decisions to limit free speech, the conditions under which banning decisions gain legitimacy. Second, we advance current literature (Hershcovis, 2011; Hershcovis & Bhatnagar, 2017; Hershcovis, Ogunfowora, Reich, & Christie, 2017) by showing how, in the context of SNS, the desire to protect the right to free speech leads a sizeable sub-group of participants to tolerate aggression. While people readily condemn aggression offline (Hershcovis & Bhatnagar, 2017; Hershcovis et al., 2017), we find that the condemnation of online aggression is limited because people still feel it is important to protect the right to free speech. Third, consistent with the predictions from the deontic theory of justice (Folger, 2001), we find that online aggression causes anger toward the aggressor, especially when the victim is high in warmth. This evidence suggests that SNS can introduce a reputation system that tracks aggression (Hanson, Jiang, & Dahl, 2019; Kim, Moravec, & Dennis, 2019; Wang, Doong, & Foxall, 2010) in an effort to increase transparency around the factors leading to the ban of certain users.

Conceptual development

Free speech controversies on SNS

Debates on SNS concerning important social and political topics can reveal impactful for several reasons. Online exchanges promote democratic forms of deliberation (Halpern & Gibbs, 2013). Likewise, the engagement on SNS supports offline volunteering and campaigning on social causes (Vaccari et al., 2015). There is also evidence that SNS expose people to a broader range of points of view (Bakshy, Messing, & Adamic, 2015).

These positive outcomes notwithstanding, scholars have noticed that the presentation of controversial social and political topics on social media is often characterized by an aggressive style and very heated interactions among users (Antoci et al., 2019; Celik, 2018; Chatzakou et

al., 2017). Users appreciate the possibility to vent their anger on SNS (Stephens, Trawley, & Ohtsuka, 2016). While this trend appears so generalized as to lead to render online aggression banal (Soral et al., 2018), there is evidence that online aggression has negative consequences on the individual user and on the online community overall. On an individual level, online aggression causes negative emotions (Bacile, Wolter, Allen, & Xu, 2018) and has negative mental health implications (Chen, 2015). For the wider community, online aggression is negative because it fosters division and lack of trust (Antoci et al., 2019). It tends to activate a small niche of partisan campaigners (Gervais, 2019) while provoking dislike and disengagement in the majority of users (Anderson et al., 2018; Barnidge, 2017).

This situation creates an imperative for SNS to limit online aggression. The temporary or permanent ban of users who contravene established rules of conduct is a necessary tool for SNS to promote constructive behavior online. Still, bans are seldom uncontroversial (Etzioni, 2019). Expelling someone permanently from a forum of debate can be construed as a way of choking or controlling public debate on controversial issues (Smith, 2018). There have been several instances of criticism by politicians and commentators claiming that SNS unduly limit free speech in order to promote a specific political agenda (Klein, 2018; Malik, 2018; Wells, 2019). Managers therefore strive to understand how users react to limitations of free speech and under which conditions they are likely to protest.

Observing aggression on SNS

When SNS legitimately ban a user? As discussed, SNS are attempting to tackle an evolving constellation of threatening, abusive, harmful or defamatory content (Antoci et al., 2019; Muddiman, 2017). Past studies have considered a wide range of online misbehavior, from incivility in political comments (Gervais, 2019), to profanity (DeFrank & Kahlbaugh, 2019; Muddiman & Stroud, 2017) and forms of hate speech or racial discrimination (Celik, 2018). Extending prior research (Hershcovis & Bhatnagar, 2017; Hershcovis, Ogunfowora, Reich, &

Christie, 2017), we focus on the social perception of aggression on SNS. Aggression is any high-intensity, targeted attack motivated by a desire to cause distress (Burt & Alhabash, 2018; Hershcovis & Bhatnagar, 2017). This is noticeably different from incivility, which is low in intensity and vague in intent (Hershcovis, Ogunfowora, Reich, & Christie, 2017).

Aggression offers the best context for studying users' reactions to bans from SNS. Despite being widespread on SNS (Antoci et al., 2019), incivility is in itself insufficient for limiting freedom of speech, given that it typically falls in the category of unpleasant rather than abusive or harmful speech (Etzioni, 2019). For example, existing policies of conduct on Twitter make it clear that incivility does not provide sufficient grounds for removing content (Twitter, 2020). On the contrary, aggression is more likely to be banned, given that it targets other users and intends to be distressful (Twitter, 2020). On a related point, aggression is more likely than incivility to activate user condemnation. Other users can recognize the victim of aggression, and therefore, in their consideration of the ban, they might weigh the importance of the right to free speech against the perceived suffering of the victim (Hershcovis & Bhatnagar, 2017; Hershcovis, Ogunfowora, Reich, & Christie, 2017). A ban is particularly likely if online aggression is recurrent and the attack continues over several online posts. With repeated aggression, that is an instance of aggression that extends over several online posts, the attack is perceived as more intense, and as a consequence, the negative intentions of the attacker might become conspicuous (Antonetti & Maklan, 2016; Porath, MacInnis, & Folkes, 2010).

We build on deontic justice theory (Folger, 2001) to explain how people react to aggression on SNS. Deontic justice theory suggests that people dislike observing others being mistreated (Folger, 2001), and will try to help the perceived victim while punishing the perpetrator (Hershcovis & Bhatnagar, 2017; Hershcovis, Ogunfowora, Reich, & Christie, 2017). This is because individuals have a sense of what is fair and of the obligations that should be respected in different social contexts (Folger, 2001). Evidence from interpersonal and organizational

psychology indicates the influence of normative fairness concerns and perceived social obligations on individual perceptions and behavior (Colquitt & Zipay, 2015; Hershcovis & Bhatnagar, 2017). When evaluating a certain behavior, individuals are able to quickly and instinctively judge whether the behavior is appropriate or not (Colquitt & Zipay, 2015). Accordingly, individuals are willing to punish others who are perceived as behaving unfairly even if this is somewhat costly for them (Aquino, Tripp, & Bies, 2006). In the context of online aggression, deontic justice theory will predict that users want to punish the aggressor who has behaved unfairly (Hershcovis & Bhatnagar, 2017; Hershcovis, Ogunfowora, Reich, & Christie, 2017). This desire to punish wrongdoing should reduce perceptions that the right to free speech is being violated by the ban while providing an implicit justification for the ban. Therefore, we hypothesize that:

H1: Aggression reduces perceptions that the ban violates free speech.

Research shows that consumers respond negatively to perceived unfairness in corporate conduct, defined as the perception that the company has violated an important moral norm (Antonetti & Maklan, 2016; Kähr, Nyffenegger, Krohmer, & Hoyer, 2016). Perceptions of a violation of free speech would form part of overall unfairness evaluations (Nielsen, 2018). In addition to free speech concerns, the overall perceived unfairness of a ban is influenced by considerations about whether a fair process was followed in taking the banning decision (Maxham & Netemeyer, 2002). As discussed above, deontic justice theory (Folger, 2001) predicts a negative reaction to online aggression and a desire to punish the perpetrator. This mechanism implies that aggression would reduce concerns about the unfairness of the ban (Hershcovis & Bhatnagar, 2017; Hershcovis, Ogunfowora, Reich, & Christie, 2017). Consequently, we hypothesize that:

H2: Aggression reduces perceptions that the ban is unfair.

Moreover, deontic justice theory (Folger, 2001) suggests that people feel angry towards the perpetrator of mistreatment (Porath, MacInnis, & Folkes, 2010). Anger motivates individuals to act against wrongdoers to reestablish justice (Antonetti & Maklan, 2016; Porath, MacInnis, & Folkes, 2010). In an interpersonal context, there is evidence that aggression is condemned because people want to support the victim (Hershcovis & Bhatnagar, 2017; Hershcovis, Ogunfowora, Reich, & Christie, 2017). In a similar vein, when a user is banned for an aggressive post, we expect that other users will recognize and condemn the aggression as a form of mistreatment, feel angry toward the aggressor and thus ultimately support the ban as a suitable form of punishment (Folger, 2001; Hershcovis & Bhatnagar, 2017).

Based on the preceding discussion, we hypothesize that:

H3: Aggression increases anger toward the aggressor.

H4: Anger toward the aggressor reduces perceptions that the ban violates free speech.

H5: Anger toward the aggressor reduces perceptions that the ban is unfair.

Responses to perceived free speech violations

The right to free speech broadly involves the freedom to express one's opinion freely (White & Crandall, 2017). This right covers symbolic and offensive speech, although the law sets out specific limitations in different jurisdictions (Marceau & Chen, 2016). While survey evidence shows that the right to free speech is important to citizens (Schaeffer, 2020; Smith, 2018), perceptions of what speech should really be free are malleable (Lindner & Nosek, 2009; Roussos & Dovidio, 2018; White & Crandall, 2017). People in general believe that freedom of speech should especially apply to utterances they agree with (Lindner & Nosek, 2009). Furthermore, ideas on what expressions of speech should be tolerated, even when one disagrees with them, change over time. For example, Chong and Levy (2018) have shown that tolerance for racist speech has decreased in the U.S. over the last two decades, even though tolerance

toward other forms of deplorable speech (e.g., militarism) has remained largely stable. Further, there is evidence of stable differences between groups, with liberals being more respectful of free speech generally (Chong & Levy, 2018; Lindner & Nosek, 2009).

Personal prejudice influences perceptions of violations to the right to free speech (Lindner & Nosek, 2009; Roussos & Dovidio, 2018). Prejudiced individuals often use free speech rhetoric in response to bans on hateful or racist utterances (Crandall, Miller, & White, 2018; White & Crandall, 2017). Therefore, personal beliefs strongly bias the evaluation of what should be protected by the right to free speech. We build on this insight and conceptualize the social networking site ban of a user as a potentially questionable act because it impoverishes online debate and hinders the free expression of a diverse set of views.

Scholars have examined other types of questionable behavior whose ethical standing is controversial and that are criticized by significant groups in society. For example, depending on their political orientation, stakeholders condemn corporate tax avoidance and punish companies that engage in this practice (Antonetti & Anesa, 2017). Similarly, the endorsement of a controversial celebrity can backfire, as some consumers' negative associations of the celebrity harm perceptions of the brand (Carrillat, O'Rourke, & Plourde, 2019). Scholars have pointed out how SNS have a social responsibility to promote free speech (Nielsen, 2018). As mentioned in the introduction, however, the majority of Americans thinks that SNS censure certain specific opinions because of their political bias (Smith, 2018). Consistent with this rationale, we expect that observers will scrutinize closely the decision of SNS to ban other users. Consumers feel angry toward companies that behave unfairly and are willing to retaliate against them (Antonetti & Maklan, 2016; Xie & Bagozzi, 2019). Punishing unfair corporate behavior is both a way of harming a company for its wrongdoing out of a desire for revenge and an attempt at forcing the company to improve its practices so that the same problem does not recur (Kähr et al., 2016; Romani, Grappi, & Bagozzi, 2013).

We focus on intentions to engage in negative word of mouth against the social networking site as a proxy for users' intentions to protest a ban they perceive as unfair and as violating the right to free speech (Antonetti & Maklan, 2016; Kähr et al., 2016). Negative word of mouth is problematic for SNS because people rely heavily on information from peers and consider such information as relatively trustworthy (Berger, 2014). Research shows that morally questionable behaviors often spark protests in the form of negative word mouth (Antonetti & Maklan, 2016), petitions (Yuksel, Thai, & Lee, 2020) and boycotts (Hawkins, 2019). In this vein, we expect that user bans considered as unjust and as violating the right to free speech will be opposed through negative word of mouth. Consequently, we hypothesize that:

H6: Perceived free speech violation increases perceived unfairness of the ban.

H7: Perceived unfairness of the ban increases negative word of mouth in protest of the ban.

H8: Perceived free speech violation increases negative word of mouth in protest of the ban.

Social identity bias in perceptions of online aggression on SNS

SNS cater for humans' need to identify with important social groups (Flanagin, Hocevar, & Samahito, 2014). When interacting on SNS, individuals engage in social identification processes, which involve the construction and expression of relevant social identities (Phillips & Broderick, 2014). Social identities have important consequences on users' perceptions and behaviors. Specifically, research shows that individuals are more concerned about the mistreatment of in-group members and are more lenient when evaluating in-group aggressors (Brewer, 2007; Haslam & Ellemers, 2005). This in-group bias is expected to play an important role in explaining how users react to observed aggression on SNS. On SNS, debates about controversial topics typically develop along partisan lines (Antoci et al., 2019; Gervais, 2019). This means that users are routinely exposed to in-group members criticizing out-group targets

and/or their ideas. In such a context, social identification processes are likely to bias users' reactions.

Research shows that individuals are biased and are more likely to think that bans of opinions they agree with represent a violation of free speech (Chong & Levy, 2019; Lindner & Nosek, 2009). This bias is expected to influence how aggression on SNS is evaluated. In a partisan debate, when an in-group member aggressively voices opinions against someone belonging to an out-group, in-group members might be willing to evaluate the aggression more leniently (Crandall, Miller, & White, 2018; Gervais, 2019). In such circumstances, the banning of an aggressive post from an in-group member will be perceived to violate the right to free speech.

In-group bias might also lead observers to minimize the negative consequences of aggression on the victim (Brewer, 2007; Haslam & Ellemers, 2005). Since perceived severity is a key antecedent of anger (Antonetti & Maklan, 2016), we expect that observers show biased perceptions about the severity of aggression, and consequently, lower anger toward the aggressor who is part of the in-group. Consistent with the in-group bias, the ban will be perceived as highly unfair if the aggressor being banned belongs to the in-group. An in-group aggressor elicits lower anger among in-group users and a higher perception that the ban violates free speech. These mechanisms would jointly increase the perception that the ban is unfair. In addition, banning an in-group member would heighten concerns about the procedural fairness of social networking sites' decisions (Maxham & Netemeyer, 2002). The heightened concern toward a member of the in-group might lead observers to question whether the social networking site implemented adequate procedures to enforce the ban. Accordingly, we hypothesize that:

H9a: The social identity of the aggressor moderates the relationship between aggression and perceptions that the ban violates free speech so that aggression reduces perceived free speech violation less (more) when the aggressor is a member of the in-group (out-group).

H9b: The social identity of the aggressor moderates the relationship between aggression and anger toward the aggressor so that aggression increases anger less (more) when the aggressor is a member of the in-group (out-group).

H9c: The social identity of the aggressor moderates the relationship between aggression and perceptions that the ban is unfair so that aggression reduces perceived unfairness of the ban less (more) when the aggressor is a member of the in-group (out-group).

A second, complementary form of in-group bias is possible. In addition to a biased perception of aggression, users might be biased in their decision to oppose the ban. In other words, it is possible that, irrespective of how the aggression episode and the ban are appraised, individuals will be less likely to protest against a social networking site when the ban targets someone from the out-group with whom they disagree. There are two reasons for this. First, users might be less motivated to put effort into supporting an out-group member (Brewer, 2007; Haslam & Ellemers, 2005). To the extent that negative word of mouth requires a voluntary and motivated action, users might not be willing to make the effort to support an out-group member they disagree with (Haslam & Ellemers, 2005; Lange & Washburn, 2012). Second, users might not perceive protesting as their responsibility. In the specific context of a partisan debate, where the differences between in-group and out-group are heightened, protesting against the ban of an opponent implies giving a platform to an opinion you disagree with. Users do not want to protest against a ban that targets an out-group member because they do not want to endorse, even if indirectly, a point of view they disagree with (Gervais, 2019; Lindner & Nosek, 2009). Based on the above, we hypothesize that:

H9d: The social identity of the aggressor moderates the relationship between perceptions that the ban is unfair and negative word of mouth so that perceived unfairness of the ban increases negative word of mouth more (less) when the aggressor is a member of the in-group (out-group).

H9e: The social identity of the aggressor moderates the relationship between perceptions that the ban violates free speech and negative word of mouth so that perceived free speech violation increases negative word of mouth more (less) when the aggressor is a member of the in-group (out-group).

Perceptions of a warm out-group victim

The preceding discussion suggests that reactions to a ban are likely to be dictated by partisan alignment. Such a perspective is in line with anecdotal evidence showing that protests in the media against the ban of high-profile users typically reflect partisan lines (e.g., Klein, 2018). As increasing polarization in SNS is widely seen as a contributing factor in the coarsening of public debate (Antoci et al., 2019), we consider a boundary condition that we postulate might be able to weaken the perverse effects of in-group bias. Accordingly, we theorize that when the victim of aggression is in the out-group and is perceived as warm, observers will be more willing to defend him or her and thus to react more positively toward the ban of an in-group member.

The literature defines warmth as a fundamental dimension that shapes social perceptions (Fiske, Cuddy, Glick, & Xu, 2002). When a social entity (e.g., individuals, social groups, nations) is perceived as warm, observers feel that the entity is fundamentally well intentioned, caring and friendly. Warmth judgments are often automatic (Fiske, Cuddy, & Glick, 2007) and have a significant impact on emotions and behaviors toward the target (Cuddy, Fiske, & Glick, 2007). High warmth groups are more likely to elicit admiration and, when attacked, feelings of compassion from others (Cuddy, Fiske, & Glick, 2007). This leads to supportive behaviors towards high-warmth groups in the form of help, a desire to cooperate and being associated with them (Ivens, Leischnig, Muller, & Valta, 2015).

In general, out-group members are perceived to be lower in warmth than in-group members (Brewer, 2007; Haslam & Ellemers, 2005). There are, however, systematic differences between how out-groups are perceived in terms of their relative warmth, and people will have a higher propensity to help out-group members who are perceived as high in warmth (Cuddy, Fiske, & Glick, 2007). In the context of our study, we expect that observers will be less likely to consider the ban from the SNS as unfair if the act of aggression targets a high-warmth out-group member (Cuddy, Fiske, & Glick, 2007). An aggression is, therefore, highly condemned when there is evidence of the likeability of the victim (Tarrant, Calitri, & Weston, 2012). In this respect, warmth is a signal that the victim is undeserving of mistreatment and should be protected (Cuddy, Fiske, & Glick, 2007). Specifically, aggression of a high-warmth victim will be perceived as a more severe violation by observers, and thus will elicit strong feelings of anger (Antonetti & Maklan, 2016). Intensified feelings of anger will then drive support for the ban. We therefore predict that:

H10: The warmth of the victim moderates the relationship between aggression and anger directed toward the aggressor so that aggression increases anger more (less) when the victim is perceived as high (low) in warmth.

Overview of the empirical research

The research model presented in Figure 1 summarizes the preceding discussion. The model theorizes a set of relationships between perceptions of online aggression, perceptions toward the aggressor, and reactions to the ban. In Figure 1, for each hypothesis outlined, we indicate the predicted (positive or negative) effects. In Study 1, we examine H1 to H9. In Study 2, we replicate the test of H1 to H8 and examine H10.

INSERT FIGURE 1 HERE

Study 1

Method

Research design and sample. We conducted a 3 (aggression: control vs. aggression vs repeated aggression) x 2 (aggressor identity: out-group vs. in-group), between-subjects, scenario-based experiment. We examined two different levels of aggression. In one condition, we consider one single, aggressive online post, while in another, we consider a case of repeated aggression, where the attack continues over several online posts. Since this is the first study examining reactions to the ban of a user, it is interesting to explore whether the hypothesized effects for aggression hold when aggression is repeated and therefore its intensity increases.

We recruited 295 American participants from the online panel Prolific Academic (Peer, Brandimarte, Samat, & Acquisti, 2017). All participants reported being Twitter users and Democratic voters at the time of the study. We first asked participants to answer questions about their political identification. Next, we presented them with a scenario depicting a tweet from a journalist (i.e., the aggressor) by the fictitious name of John M. Kenyon, who commented on a new environmental plan introduced by a (fictitious) representative, Richard Matley. In the out-group aggressor (in-group aggressor) condition, a conservative (liberal) journalist attacked the representative. Next, participants read that Twitter had banned the aggressor and answered questions measuring perceived free speech violation, perceived unfairness of the ban, anger toward the aggressor and intentions to engage in negative word of mouth to protest the ban. The survey lasted 10 minutes and participants received $1 for their participation. The sample included 54% female participants and different age groups: 29% were 18 to 24 years old, 33% were 25 to 34 years old, 22% were 35 to 44 years old, 10% were 45 to 55 years old, and 6% were 55 years of age or older.

Stimuli. We developed the scenarios based on the literature (Gervais, 2019) and extensive secondary research seeking to emulate exchanges of real-life tweets. We also pre-tested the scenarios with a similar sample of participants (N = 299). As we wanted to explore reactions to a ban, our aggression and repeated aggression conditions included different forms of communication that in the literature have been associated with mistreatment. First, we included rude language and the use of epithets (Gervais, 2019). Second, we included profanity that suggests an intense aggression and is often disliked by audiences (DeFrank & Kahlbaugh, 2019; Muddiman & Stroud, 2017). Third, we included aggression directed at a specific individual (i.e. the representative). Direct aggression is problematic in terms of its consequences for the victim and because it contradicts SNS rules of conduct. Moreover, the repeated aggression condition, where the aggression extended over three tweets, included a wider range of aggressive terms. Given the debate topic and our focus on examining a partisan context, we changed the nature of the opposition to the environmental plan across the tweets. When a Democratic representative proposed the plan, the conservative pundit opposed it for its restrictions on business. By contrast, when a Republican representative proposed the environmental plan, the liberal pundit opposed it for its negative impact on the environment (see detailed scenarios in Appendix A).

As manipulation checks, we used items measuring perceptions of aggressiveness (e.g., "John M. Kenyon insulted representative Matley" – rated from 1 = strongly disagree to 7 = strongly agree). Participants in the aggression and repeated aggression conditions felt that John M. Kenyon "made inappropriate remarks to representative Matley" ($M_{control}$ = 3.18, $M_{aggress}$ = 5.58, $M_{repeated\_aggress}$ = 6.02; F (2, 294) = 96.31, $p < .001$), "attacked representative Matley in an unacceptable manner" ($M_{control}$ = 3.14, $M_{aggress}$ = 5.39, $M_{repeated\_aggress}$ = 6.01; F (2, 294) = 98.29, $p < .001$), "insulted representative Matley" ($M_{control}$ = 3.82, $M_{aggress}$ = 5.94, $M_{repeated\_aggress}$ = 6.28; F (2, 294) = 94.38, $p < .001$), "was rude to representative Matley" ($M_{control}$ = 3.92, $M_{aggress}$ =

6.05, $M_{repeated\_aggress}$ = 6.27; F (2, 294) = 83.90, $p < .001$), "was nasty to representative Matley" ($M_{control}$ = 3.47, $M_{aggress}$ = 5.91, $M_{repeated\_aggress}$ = 6.31; F (2, 294) = 132.63, $p < .001$), and "was disagreeable" ($M_{control}$ = 4.97, $M_{aggress}$ = 5.92, $M_{repeated\_aggress}$ = 6.18; F (2, 294) = 25.65, $p < .001$). On average, the aggression manipulation was successful (F (2, 294) = 120.61, $p < .001$). We did find significant differences between the control and aggression conditions ($M_{control}$ = 3.75, $M_{aggress}$ = 5.80; t(182) = 11.74, $p < .001$), as well as between aggression and repeated aggression conditions ($M_{aggress}$ = 5.80, $M_{repeated\_aggress}$ = 6.18; t(195) = 2.50, $p < .05$).

At the end of the experiment, we checked the manipulation of aggressor identity by asking participants to rate their perceived similarity with the journalist based on four items. Participants in the Democrat pundit condition rated John M. Kenyon as; "very close to them" ($M_{Democrat}$ = 2.41, $M_{Republican}$ = 1.70, t(293) = 4.83, $p < .001$), as "belonging to the same in-group" ($M_{Democrat}$ = 3.57, $M_{Republican}$ = 1.87, t(293) = 9.85, $p < .001$), as "very similar to them" ($M_{Democrat}$ = 3.01, $M_{Republican}$ = 1.77, t(293) = 7.89, $p < .001$), and "just like them" ($M_{Democrat}$ = 2.62, $M_{Republican}$ = 1.78, t(293) = 5.48, $p < .001$). Overall, the ratings were in line with our expectations ($M_{Democrat}$ = 2.90, $M_{Republican}$ = 1.78, t(293) = 7.82, $p < .001$). Finally, participants evaluated the realism of the tweets through four items rated from 1 = strongly disagree to 7 = strongly agree. They confirmed that the tweets were clear (M=5.68), realistic (M = 5.50), believable (M = 5.43), and representative of tweets encountered on Twitter before the research (M = 5.15). No differences in realism ratings were found across conditions ($p > .05$).

Measures. We borrowed scales from prior literature (see Appendix B for details on the items). We measured perceived unfairness of the ban with three items from Antonetti and Maklan (2016), perceived free speech violation with four items from White and Crandall (2017), anger toward the aggressor with three items from Porath et al. (2010), and negative word of mouth intentions with three items from Antonetti and Maklan (2016). Appendix B shows that all scales were reliable, with high loadings on the intended constructs. Average

Variance Extracted (AVE) and Composite Reliability (CR) were above the established thresholds. The Fornell-Larcker criterion (Fornell & Larcker, 1981) confirmed discriminant validity.

Results

We conducted a MANOVA with aggression and aggressor identity as independent variables. Results show a significant main effect of aggression on perceived free speech violation ($M_{control}$ = 5.44, $M_{aggress}$ = 4.34, $M_{repeated\_aggress}$ = 3.86; F (2, 294) = 23.26, $p < .001$), perceived unfairness of the ban ($M_{control}$ = 5.14, $M_{aggress}$ = 3.74, $M_{repeated\_aggress}$ = 3.32; F (2, 294) = 34.13, $p < .001$), negative word of mouth ($M_{control}$ = 3.86, $M_{aggress}$ = 2.62, $M_{repeated\_aggress}$ = 2.43; F (2, 294) = 24.86, $p < .001$), and anger toward the aggressor ($M_{control}$ = 3.83, $M_{aggress}$ = 4.37, $M_{repeated\_aggress}$ = 4.65; F (2, 294) = 6.65, $p < .001$). Likewise, there is a main effect of aggressor identity on perceived unfairness of the ban ($M_{Democrat}$ = 4.31, $M_{Republican}$ = 3.83, t(293) = 2.34, $p < .05$), negative word of mouth ($M_{Democrat}$ = 3.37, $M_{Republican}$ = 2.58, t(293) = 4.23, $p < .001$) and anger toward the aggressor ($M_{Democrat}$ = 3.87, $M_{Republican}$ = 4.70, t(293) = 4.55, $p < .001$), while the effect on free speech violation is not significant ($M_{Democrat}$ = 4.65, $M_{Republican}$ = 4.44, t(293) = 1.03, $p > .05$).

Further, we find a significant interaction between aggression and aggressor identity on negative word of mouth (F (2, 287) = 6.30, $p < .05$). Negative word of mouth against the ban is greater when there is no aggression (vs. aggression or repeated aggression) and the journalist is part of the in-group (i.e., a liberal pundit) ($M_{control}$ = 4.71, $M_{aggression}$ = 2.89, $M_{repeated\_aggress}$ = 2.55; F (2, 145) = 26.45, $p < .001$). In other words, when there is no aggression, users oppose a ban against a journalist who belongs to the in-group. Opposition of the ban is not as high if there is aggression involved, and is even less so if aggression is repeated over time. By contrast, when the journalist belongs to the out-group, negative word mouth against the ban is always low ($M_{control}$ = 3.06, $M_{aggression}$ = 2.36, $M_{repeated\_aggress}$ = 2.31; F (2, 148) = 4.63, $p < .05$). Figure

2 illustrates the interaction effect. No interaction effect between aggression and aggressor identity is found when considering free speech violation (p = .594), perceived unfairness of the ban (p = .526) and anger toward the aggressor (p = .394). Thus, H9a, H9b and H9c are not supported by the data.

INSERT FIGURE 2 HERE

An interesting descriptive finding concerns also the relative acceptability of aggression on the grounds of the right to free speech. When the aggressor belongs to the out-group, 51% of participants in the aggression condition and 40% in the repeated aggression condition scored above four on the measure of perceived free speech violation. When the aggressor belongs to the in-group, the values are even higher, reaching 59% and 43%, respectively. This evidence suggests that, while aggression reduces perceptions that the ban violates free speech, a significant segment of observers considers aggression broadly justifiable in light of their desire to protect the right to free speech.

To test our moderated mediation model, we conducted a regression-based conditional effect analysis using a custom model in PROCESS (Hayes, 2018). The model was estimated using 10,000 resamples for the calculation of confidence intervals (CIs) and used bias-corrected and accelerated bootstrap (Hayes, 2018). Given that the factor of aggression was set at two levels – aggression and repeated aggression – we estimated the model in two consecutive runs; first by comparing control vs aggression conditions, and second by comparing control vs repeated aggression conditions. We included age and gender as covariates. Results are presented in Table 1[1]. As hypothesized, we find two significant interaction effects on negative word of mouth; one between free speech violation and aggressor's identity and another between perceived unfairness of the ban and aggressor's identity. The effect of free speech on negative word of

[1] The aggression variable was coded as -1 control and +1 aggression. Aggressor identity was coded as -1 in case of out-group and +1 in case of in-group. We used the average of all the items for the analysis. Since H9a, H9b, and H9c are not supported by the data, we report the results here without the estimation of these hypotheses.

mouth depends on whether the aggressor is part of the in-group (vs. out-group) and so does the effect of perceived unfairness of the ban. The analysis of conditional indirect effects is presented in Table 2. We find that the mediation of perceived unfairness of the ban, and the serial mediation of perceived free speech violation and perceived unfairness of the ban are supported when the ban targets an in-group aggressor.

INSERT TABLE 1 AND TABLE 2 HERE

Discussion

The results from Study 1 largely support our moderated mediation model. Users oppose the banning decision on the grounds of free speech violation and perceptions of injustice/unfairness. As a form of opposition, users are willing to spread negative word of mouth to protest the ban, consistent with H6-H8. Notably, we find that the increase in negative word of mouth is particularly evident if the ban affects a user who is part of the in-group, thus shares similar political views. This evidence offers support for H9d. When online aggression is present, observers report anger toward the aggressor and lower perceptions of free speech violation and unfairness, in line with H1, H2 and H3.

Nonetheless, Study 1 does not elucidate factors that might increase users' acceptance of the decision of the social networking site to ban an in-group member. We hypothesize that users' acceptance of a ban increases when the out-group victim is perceived as warm (H10). This hypothesis is examined in Study 2.

Study 2

Method

Research design and sample. We conducted a 2 (aggression: control vs. aggression) X 2 (warmth of the out-group victim: low vs. high) between-subjects experiment. We dropped the repeated aggression condition as Study 1 shows that there are no significant differences between the single aggression and the repeated aggression conditions. We recruited 297 Twitter users

through Prolific Academic, following the procedures described in Study 1. All participants were Democratic voters. After indicating their political identification, participants read a tweet from the victim of aggression commenting on an environmental plan introduced by the Republicans. Subsequently, they read a response from a fictitious liberal journalist John M. Kenyon (i.e., the aggressor). Next, the decision to ban the journalist was introduced, and participants answered a series of questions. Participants were 47% female and represented different age groups: 18% were 18-24 years old, 37% 25-34, 21% 35-44, 13% 45-55, and 11% 55 or above.

Stimuli. Participants were first informed of the introduction, by Republicans, of an environmental plan aimed at reducing environmental constraints on businesses. Next, they read a tweet supportive of the plan. In the high-warmth condition, the tweet was from Richard Matley, a Professor of Environmental Science acting as *"volunteer for several environmental protection charities providing free scientific advice on a number of environmental issues."* In the low-warmth condition, Richard Matley was presented as a Professor of Environmental Science acting as *"paid consultant to the boards of several gas and oil corporations and to their lobbies, providing scientific advice on a number of environmental issues."* We manipulated aggression using a text similar to the one employed in Study 1 (see details in Appendix C). Participants perceived the aggressive tweets as clear (M = 5.75), realistic (M = 5.65), believable (M = 5.61) and representative of tweets encountered before the research (M = 4.68), with no significant differences between conditions ($p > .05$).

We successfully manipulated aggression ($M_{control} = 3.77$, $M_{aggress} = 5.15$; $t(295) = 11.62$, $p < .001$). We also checked the manipulation of warmth of the victim by asking participants to rate perceived warmth based on four items from Fiske, Cuddy, Glick, and Xu (2002; see also Liu, Bogicevic, & Mattila, 2018). The warmth ratings were in line with our expectations ($M_{low\_warmth} = 3.68$, $M_{high\_warmth} = 4.07$, $t(295) = 3.09$, $p < .05$).

Measures. We retained all the scales used in Study 1. All scales performed satisfactorily in terms of reliability, as demonstrated by the standardized loadings, CR and AVE reported in Appendix B. There was also evidence of acceptable discriminant validity (Fornell & Larcker, 1981).

Results

A MANOVA showed a main effect of aggression on perceived free speech violation ($M_{control}$ = 5.23, $M_{aggress}$ = 4.60; F (1, 291) = 12.65, p < .01), anger toward the aggressor ($M_{control}$ = 2.99, $M_{aggress}$ = 3.78; F (1, 291) = 21.37, p < .01), perceived unfairness of the ban ($M_{control}$ = 5.27, $M_{aggress}$ = 4.37; F (1, 291) = 26.77, p < .01), and negative word of mouth ($M_{control}$ = 4.29, $M_{aggress}$ = 3.55; F (1, 291) = 15.33, p < .01). Perceived warmth of the victim influenced anger toward the aggressor (M_{low_warmth} = 3.09, M_{high_warmth} = 3.67; F (1, 291) = 12.24, p < .05) and perceived unfairness of the ban (M_{low_warmth} = 5.02, M_{high_warmth} = 4.61; F (1, 291) = 4.58, p < .05). We also found an interaction effect of aggression and warmth of the victim on anger toward the aggressor (F (1, 291) = 5.05, p < .05). Figure 3 illustrates the interaction effect on anger toward the aggressor. No other interaction effects were statistically significant.

As in Study 1, we conducted a conditional effect analysis to test the moderated mediation model presented in Figure 1. Table 1 reports the path estimates[2]. Consistent with the results from Study 1, we found that aggression increases anger toward the aggressor, while diminishing perceptions of free speech violation and perceived unfairness of the ban. Both free speech violation and perceived unfairness of the ban are drivers of negative word of mouth. Notably, the relationship between aggression and anger toward the aggressor is moderated by warmth of the victim. Aggression has a positive effect on anger when the victim is high in warmth, but the effect is not significant when the victim is low in warmth.

[2] The warmth variable was coded as -1 low warmth and +1 high warmth. The aggression condition was coded as -1 control and +1 aggression. We used the average of all the items for the analysis.

Table 3 reports the conditional indirect effects analysis. The indirect effect of aggression on negative word of mouth, mediated by anger toward the aggressor, free speech violation and perceived unfairness of the ban, is statistically significant only when the victim is high in warmth.

INSERT FIGURE 3 AND TABLE 3 HERE

Discussion

Consistent with Study 1, we find that online aggression increases anger toward the aggressor while diminishing perceptions of free speech violation and unfairness. The study also replicates the findings of Study 1 and offers support to H1-H8. Further, we find support for H10. Aggression increases anger toward the aggressor when the victim does not share similar political views, but is high in warmth. This leads to increased acceptance of the ban as a way to punish the aggressor.

General discussion

Theoretical implications

Consistent with the objectives of the Special Issue, the study raises important implications for research on the psychological factors that influence stakeholders' reactions to the protection of free speech. While past research demonstrates that observers proactively condemn aggression (Anderson et al., 2018; Barnidge, 2017), this study is the first to examine the conditions under which users of SNS welcome or even condemn bans aimed at suppressing online aggression. We put forward the view that observers of online aggression evaluate the ban in terms of perceived (un)fairness, which results from the trade-off between protecting the right to free speech and suppressing online aggression. Our findings show that, although observers dislike aggression (Hershcovis & Bhatnagar, 2017), bans are carefully scrutinized by users and can be considered as questionable or controversial and thus cause reputational fallouts

(Antonetti & Anesa, 2017; Carrillat, O'Rourke, & Plourde, 2019; Xie & Bagozzi, 2019). Perceptions that the ban of a SNS user violates free speech and is unfair motivate observers to protest against a social networking site by engaging in negative word of mouth. At the same time, however, we provide evidence of an in-group bias (Brewer, 2007; Haslam & Ellemers, 2005) regarding the motivation to protest against the ban of a user. In contrast with normative arguments on the importance of free speech (Etzioni, 2019; Nielsen, 2018), users are willing to protest against limitations to free speech only when these affect members of the in-group. This is an important finding advancing understanding of the psychology of free speech: people are not willing to defend free speech universally, but only when they agree with the opinions being suppressed in the first place.

A novel contribution of this research also lies in the application of the deontic theory of justice (Folger, 2001) to the analysis of the effect of perceived aggression on users' reactions to a ban. We find that aggression is relevant to the evaluation of a social networking site ban because it reduces perceptions that the ban violates free speech and is unfair. These effects extend the relevance of deontic justice theory to different forms of mistreatment, perpetrated online (Hershcovis, 2011; Hershcovis & Bhatnagar, 2017; Hershcovis, Ogunfowora, Reich, & Christie, 2017). We also contribute to debates on the social perception of online aggression (Antoci et al., 2019; Hershcovis, 2011; Muddiman, 2017) by demonstrating that even repeated aggression is tolerated by a sizeable subgroup of participants on the grounds of free speech. This evidence stresses how complex it is for users to balance free speech concerns with the need to suppress online aggression. It is worth noting that, in episodes of repeated online aggression where the attack is severe and protracted over time, a large segment of observers does not appear to accept the ban due to a strong desire to protect the right to free speech. This observation also highlights an interesting difference in perceptions of offline and online aggression. Free speech does not play a role in the evaluation of aggression occurring in offline social settings

(Hershcovis, 2011). The finding implies that reducing online aggression might be intrinsically more challenging.

Nonetheless, a consistent result across both studies is that aggression causes anger toward the aggressor, even when the aggressor belongs to the in-group and thus shares the political views of the observers of the ban. This echoes further evidence on reactions to online aggression (Anderson et al., 2018; Barnidge, 2017), and suggests that observers judge an aggressive user negatively, irrespective of whether they might act to punish such aggression. Theoretically, this insight suggests the importance of studying further how aggression influences users' reputations on SNS. Scholars have examined the importance and value of reputation systems on SNS (Hanson, Jiang, & Dahl, 2019; Kim, Moravec, & Dennis, 2019; Wang, Doong, & Foxall, 2010). Yet, to date, existing reputation systems do not appear to account for the level of aggression that users display. Managerially, as we examine next, this finding offers an incentive to SNS to introduce transparent reputation systems based on user aggression.

Managerial and policy implications

Our studies offer notable implications for managers of SNS and insights for policy makers. First, our findings indicate that users closely scrutinize how SNS handle the controversies arising from political debates. SNS users seem to generally oppose bans and leverage the rhetoric of free speech to justify their opposition to these restrictions. Limitations to free speech represent a controversial or ethically questionable behavior that users are willing to punish (Antonetti & Anesa, 2017; Carrillat, O'Rourke, & Plourde, 2019; Xie & Bagozzi, 2019). SNS users resort to negative word of mouth as a form of protest against the ban. In light of the above evidence, we advise SNS to evaluate cautiously decisions concerning the ban of SNS users. In circumstances where the ban is inevitable because of the high levels of online aggression, SNS should justify their decision to other users observing the ban. In particular, we recommend managers to provide a detailed explanation of the reasons why the ban should not be interpreted

as a limitation to users' right to free speech, but rather as a mechanism toward ensuring fair treatment of SNS users. Another critical implication, highlighted in recent coverage of Twitter's policy on the ban of aggressive content (Walawalkar, 2020), concerns the need to show consistency in decisions to ban users and/or content. Inconsistent decisions or double-standards are likely to increase perceived unfairness and lead to user protests against bans.

Second, we show that aggression reduces the likelihood that users might protest a ban. Aggression elicits anger toward the aggressor, especially when the attacked victim is known to be warm. Our findings indicate, therefore, that users might support bans framed as a proactive step taken to address online aggression. SNS should therefore consider framing explicitly bans as necessary actions aimed at ensuring free treatment of SNS users, and at helping the victims of online aggression. Yet, this implication comes with a note of caution to SNS. While aggression increases the acceptability of a ban, free speech concerns still remain important. Explicitly stating the aim of SNS to suppress online aggression could serve as a meaningful justification for the ban, yet this might not always be sufficient. In this respect, SNS could consider introducing reputation management mechanisms that track and penalize aggression in a transparent and consistent manner. This could translate into the development of a reputation score capturing users' aggression levels on SNS. Artificial Intelligence could help tracking online aggression and enabling real-time updates of the reputation scores (Marr, 2019; Simonite, 2020). Efforts in this direction would convey enhanced transparency of the decision-making process leading to the ban of an aggressive user.

More broadly, our research elucidates the complex trade-offs underlying the reactions to SNS user bans. Evidence from our studies show that SNS users are concerned about justice and the punishment of online aggression as much as they are concerned about preserving their right to free speech. The implication for both SNS and policy makers concerns the need to institute much broader and more transparent discussions on what kind of speech can be restricted on

SNS, especially when debating political issues of wide societal relevance. Arguably, SNS do not have the legitimacy to decide independently on such a sensitive issue. Policy makers, in this respect, are starting to introduce regulations that address hate speech and aggression, yet still protect the right to free speech (e.g., Breeden, 2020). It is advisable for policy makers and SNS to work closely to define acceptable and sensible guidelines that will help regulate the online environment. SNS users should also be consulted in such a process in order to make sure that there is a high level of acceptance for any change introduced. We recommend bringing together policy makers, SNS and users to co-create regulations and codes of practice. The current environment of uncertainty, where policies are scant or inconsistently applied, leads to bans often being challenged and such a practice risks heightening users' skepticism in the long run (Antoci et al., 2019). Policy makers and SNS should ultimately consider structural changes that promote civil discussions on SNS while allowing users to exercise their right to free speech.

Limitations and areas for further research

While offering the first systematic examination of how observers react to SNS bans, the study presents some limitations that provide interesting avenues for further research. The first limitation concerns the fact that our study focuses on a specific type of ban not necessarily representative of all conditions where SNS might decide to ban users, and restrict the right to free speech. Specifically, the studies presented focus on the research context of environmental policy debates, consider political identities to explore in-group and out-group dynamics and focus specifically on bans affecting users rather than considering bans targeting content only. We focus on this research context because of its practical relevance in advancing timely debates around the role of SNS in limiting the right to free speech (DuMont, 2016; Etzioni, 2019; Nielsen, 2018). Nonetheless, future research would benefit from examining other types of bans and other research contexts.

Furthermore, while building on deontic justice theory (Folger, 2001), we only consider the support of a ban as a way to help the victim of aggression. This choice is justified by our interest in how observers react to a ban that restricts freedom of speech. Nonetheless, it is possible to expand the analysis significantly, as observers on SNS might act to support a victim directly (Hershcovis & Bhatnagar, 2017; Hershcovis, Ogunfowora, Reich, & Christie, 2017). Future research can examine observers' decisions to either attack the aggressor or to report the aggressor to the SNS as two other important outcomes that might influence the evolution of aggressive exchanges on SNS.

Finally, this research focuses on aggression as one specific form of online mistreatment. There are, however, other forms of questionable behavior that call for limitations to the right of freedom of speech. Online bullying (Chatzakou et al., 2017) and the spread of false information (Kim, Moravec, & Dennis, 2019) might, arguably, lead to the banning of specific posts and sometimes of certain users. These two other forms of online misbehavior have sparked complex debates about the responsibility of SNS to protect users while also respecting the right to free speech (Etzioni, 2019; Marceau & Chen, 2016). Future research should consider other types of online misbehavior to examine how observers evaluate the trade-off between imposing restrictions to certain users while protecting the right to free speech.

References

Anderson, A. A., Yeo, S. K., Brossard, D., Scheufele, D. A., & Xenos, M. A. (2018). Toxic talk: How online incivility can undermine perceptions of media. *International Journal of Public Opinion Research*, *30*(1), 156-168.

Antoci, A., Bonelli, L., Paglieri, F., Reggiani, T., & Sabatini, F. (2019). Civility and trust in social media. *Journal of Economic Behavior & Organization*, *160*, 83-99.

Antonetti, P., & Anesa, M. (2017). Consumer reactions to corporate tax strategies: The role of political ideology. *Journal of Business Research*, *74*, 1-10.

Antonetti, P., & Maklan, S. (2016). An extended model of moral outrage at corporate social irresponsibility. *Journal of Business Ethics*, *135*(3), 429-444.

Aquino, K., Tripp, T. M., & Bies, R. J. (2006). Getting even or moving on? Power, procedural justice, and types of offense as predictors of revenge, forgiveness, reconciliation, and avoidance in organizations. *Journal of Applied Psychology*, *91*(3), 653-668.

Bacile, T. J., Wolter, J. S., Allen, A. M., & Xu, P. (2018). The effects of online incivility and consumer-to-consumer interactional justice on complainants, observers, and service providers during social media service recovery. *Journal of Interactive Marketing*, *44*, 60-81.

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130-1132.

Barnidge, M. (2017). Exposure to political disagreement in social media versus face-to-face and anonymous online settings. *Political Communication*, *34*(2), 302-321.

BBC News (2019, October 30). Meghan Murphy: Canadian feminist's trans talk sparks uproar. Retrieved from https://www.bbc.com/news/world-us-canada-50214341.

Berger, J. (2014). Word of mouth and interpersonal communication: A review and directions for future research. *Journal of Consumer Psychology*, *24*(4), 586-607.

Breeden, A. (2020, June 18). French court strikes down most of online hate speech law. *The New York Times*. Retrieved from https://www.nytimes.com/2020/06/18/world/europe/france-internet-hate-speech-regulation.html.

Brewer, M. B. (2007). The importance of being we: Human nature and intergroup relations. *American Psychologist*, *62*(8), 728-738.

Burt, S. A., & Alhabash, S. (2018). Illuminating the nomological network of digital aggression: Results from two studies. *Aggressive Behavior*, *44*(2), 125-135.

Carrillat, F. A., O'Rourke, A. M., & Plourde, C. (2019). Celebrity endorsement in the world of luxury fashion–when controversy can be beneficial. *Journal of Marketing Management*, *35*(13-14), 1193-1213.

Celik, S. (2018). Tertiary-level internet users' opinions and perceptions of cyberhate. *Information Technology & People*, *31*(3), 845-868.

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017). Mean birds: Detecting aggression and bullying on twitter. *In Proceedings of the 2017 ACM on Web Science Conference* (pp. 13-22).

Chen, G. M. (2015). Losing face on social media: Threats to positive face lead to an indirect effect on retaliatory aggression through negative affect. *Communication Research*, *42*(6), 819-838.

Chong, D., & Levy, M. (2018). Competing Norms of Free Expression and Political Tolerance. *Social Research: An International Quarterly*, *85*(1), 197-227.

Colquitt, J. A., & Zipay, K. P. (2015). Justice, Fairness, and Employee Reactions. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*, 75-99.

Crandall, C. S., Miller, J. M., & White, M. H. II (2018). Changing norms following the 2016 US presidential election: The Trump effect on prejudice. *Social Psychological and Personality Science*, *9*(2), 186-192.

Cuddy, A. J., Fiske, S. T., & Glick, P. (2007). The BIAS map: behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology*, *92*(4), 631-648.

DeFrank, M., & Kahlbaugh, P. (2019). Language choice matters: When profanity affects how people are judged. *Journal of Language and Social Psychology*, *38*(1), 126-141.

DuMont, S. (2016). Campus Safety v. Freedom of Speech: An Evaluation of University Responses to Problematic Speech on Anonymous Social Media. *Journal of Business & Technology Law*, *11*, 239-264.

Etzioni, A. (2019). Allow Offensive Speech-Curb Abusive Speech? *Society*, *56*(4), 315-321.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77-83.

Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878-902.

Flanagin, A. J., Hocevar, K. P., & Samahito, S. N. (2014). Connecting with the user-generated Web: how group identification impacts online information sharing and evaluation. *Information, Communication & Society*, *17*(6), 683-694.

Folger, R. (2001). Fairness as deonance. In S. W. Gilliland, D. D. Steiner, & D. P. Skarlicki (Eds.), *Research in Social Issues in Management* (pp. 3–31). Charlotte, NC: Information Age.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39-50.

Gervais, B. T. (2019). Rousing the partisan combatant: Elite incivility, anger, and antideliberative attitudes. *Political Psychology*, *40*(3), 637-655.

Halliday, J. (2012, March 22). Twitter's Tony Wang: "We are the free speech wing of the free speech party". *The Guardian.* Retrieved from https://www.theguardian.com/media/2012/mar/22/twitter-tony-wang-free-speech.

Halpern, D., & Gibbs, J. (2013). Social media as a catalyst for online deliberation? Exploring the affordances of Facebook and YouTube for political expression. *Computers in Human Behavior*, *29*(3), 1159-1168.

Hanson, S., Jiang, L., & Dahl, D. (2019). Enhancing consumer engagement in an online brand community via user reputation signals: a multi-method analysis. *Journal of the Academy of Marketing Science*, *47*(2), 349-367.

Haslam, S. A., & Ellemers, N. (2005). Social identity in industrial and organizational psychology: Concepts, controversies and contributions. *International Review of Industrial and Organizational Psychology*, *20*(1), 39-118.

Hawkins, M. A. (2019). The effect of activity identity fusion on negative consumer behavior. *Psychology & Marketing*, *36*(4), 395-409.

Hayes, A. F. (2018). *Introduction to Mediation, Moderation and Conditional Process Analysis* (2nd ed). Guilford Press: New York, NY.

Hershcovis, M. S. (2011). "Incivility, social undermining, bullying… oh my!": A call to reconcile constructs within workplace aggression research. *Journal of Organizational Behavior*, *32*(3), 499-519.

Hershcovis, M. S., & Bhatnagar, N. (2017). When fellow customers behave badly: Witness reactions to employee mistreatment by customers. *Journal of Applied Psychology*, *102*(11), 1528-1544.

Hershcovis, M. S., Ogunfowora, B., Reich, T. C., & Christie, A. M. (2017). Targeted workplace incivility: The roles of belongingness, embarrassment, and power. *Journal of Organizational Behavior*, *38*(7), 1057-1075.

Ivens, B. S., Leischnig, A., Muller, B., & Valta, K. (2015). On the role of brand stereotypes in shaping consumer response toward brands: An empirical examination of direct and mediating effects of warmth and competence. *Psychology & Marketing*, *32*(8), 808-820.

Kähr, A., Nyffenegger, B., Krohmer, H., & Hoyer, W. D. (2016). When hostile consumers wreak havoc on your brand: The phenomenon of consumer brand sabotage. *Journal of Marketing*, *80*(3), 25-41.

Kim, A., Moravec, P. L., & Dennis, A. R. (2019). Combating fake news on social media with source ratings: The effects of user and expert reputation ratings. *Journal of Management Information Systems*, *36*(3), 931-968.

Klein, E. (2018, August 8). The problem with Twitter, as shown by the Sarah Jeong fracas. *Vox*. Retrieved from https://www.vox.com/technology/2018/8/8/17661368/sarah-jeong-twitter-new-york-times-andrew-sullivan.

Lange, D., & Washburn, N. T. (2012). Understanding attributions of corporate social irresponsibility. *Academy of Management Review*, *37*(2), 300-326.

Lindner, N. M., & Nosek, B. A. (2009). Alienable speech: Ideological variations in the application of free-speech principles. *Political Psychology*, *30*(1), 67-92.

Liu, S. Q., Bogicevic, V., & Mattila, A. S. (2018). Circular vs. angular servicescape: "Shaping" customer response to a fast service encounter pace. *Journal of Business Research*, *89*, 47-56.

Malik, K. (2018, November 25). Debate ends when we label views we simply disagree with as 'hatred', *The Guardian*. Retrieved from https://www.theguardian.com/commentisfree/2018/nov/25/debate-ends-when-we-label-views-we-disagree-with-us-hatred.

Marceau, J. F., & Chen, A. K. (2016). Free Speech and Democracy in the Video Age. *Columbia Law Review*, *991*, 15-42.

Marr, B. (2019, August 30). Can artificial intelligence predict the spread of online hate speech? *Forbes*. Retrieved from https://www.forbes.com/sites/bernardmarr/2019/08/30/can-artificial-intelligence-predict-the-spread-of-online-hate-speech/

Maxham I, J. G. II, & Netemeyer, R. G. (2002). A longitudinal study of complaining customers' evaluations of multiple service failures and recovery efforts. *Journal of Marketing*, *66*(4), 57-71.

Muddiman, A. (2017). Personal and public levels of political incivility. *International Journal of Communication*, *11*, 3182-3202.

Muddiman, A., & Stroud, N. J. (2017). News values, cognitive biases, and partisan incivility in comment sections. *Journal of Communication*, *67*(4), 586-609.

Nielsen, R. P. (2018). Ethical and Legal First Amendment Implications of FBI v. Apple: A Commentary on Etzioni's 'Apple: Good Business, Poor Citizen?'. *Journal of Business Ethics*, *151*(1), 17-28.

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153-163.

Phillips, N., & Broderick, A. (2014). Has Mumsnet changed me? SNS influence on identity adaptation and consumption. *Journal of Marketing Management*, *30*(9-10), 1039-1057.

Porath, C., MacInnis, D., & Folkes, V. (2010). Witnessing incivility among employees: Effects on consumer anger and negative inferences about companies. *Journal of Consumer Research*, *37*(2), 292-303.

Romani, S., Grappi, S., & Bagozzi, R. P. (2013). My anger is your gain, my contempt your loss: Explaining consumer responses to corporate wrongdoing. *Psychology & Marketing*, *30*(12), 1029-1042.

Roussos, G., & Dovidio, J. F. (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.

Schaeffer, K. (2020, May 29). Fast facts about Americans' views of social media companies as Trump-Twitter dispute grows. *Pew Research Center.* Retrieved from https://www.pewresearch.org/fact-tank/2020/05/29/fast-facts-about-americans-views-of-social-media-companies-as-trump-twitter-dispute-grows/.

Simonite, T. (2020, December 5). Facebook's AI for hate speech improves. How much is unclear. *Wired*. Retrieved from https://www.wired.com/story/facebook-ai-hate-speech-improves-unclear/.

Smith, A. (2018, June 28). Public attitudes toward technology companies. *Pew Research Center.* Retrieved from https://www.pewresearch.org/internet/2018/06/28/public-attitudes-toward-technology-companies/.

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

Stephens, A. N., Trawley, S. L., & Ohtsuka, K. (2016). Venting anger in cyberspace: Self-entitlement versus self-preservation in road rage tweets. *Transportation Research Part F: Traffic Psychology and Behaviour*, *42*(2), 400-410.

Tarrant, M., Calitri, R., & Weston, D. (2012). Social identification structures the effects of perspective taking. *Psychological Science*, *23*(9), 973-978.

Twitter (2020). *The Twitter Rules*. Retrieved from https://help.twitter.com/en/rules-and-policies/twitter-rules.

Vaccari, C., Valeriani, A., Barberá, P., Bonneau, R., Jost, J. T., Nagler, J., & Tucker, J. A. (2015). Political expression and action on social media: Exploring the relationship between lower-and higher-threshold political activities among Twitter users in Italy. *Journal of Computer-Mediated Communication*, *20*(2), 221-239.

Walawalkar, A. (2020). Twitter accused of double standards over ban on tweets wishing death on Trump. *The Observer*. Retrieved from: https://www.theguardian.com/technology/2020/oct/03/twitter-faces-backlash-over-abuse-policy-in-wake-of-trump-illness.

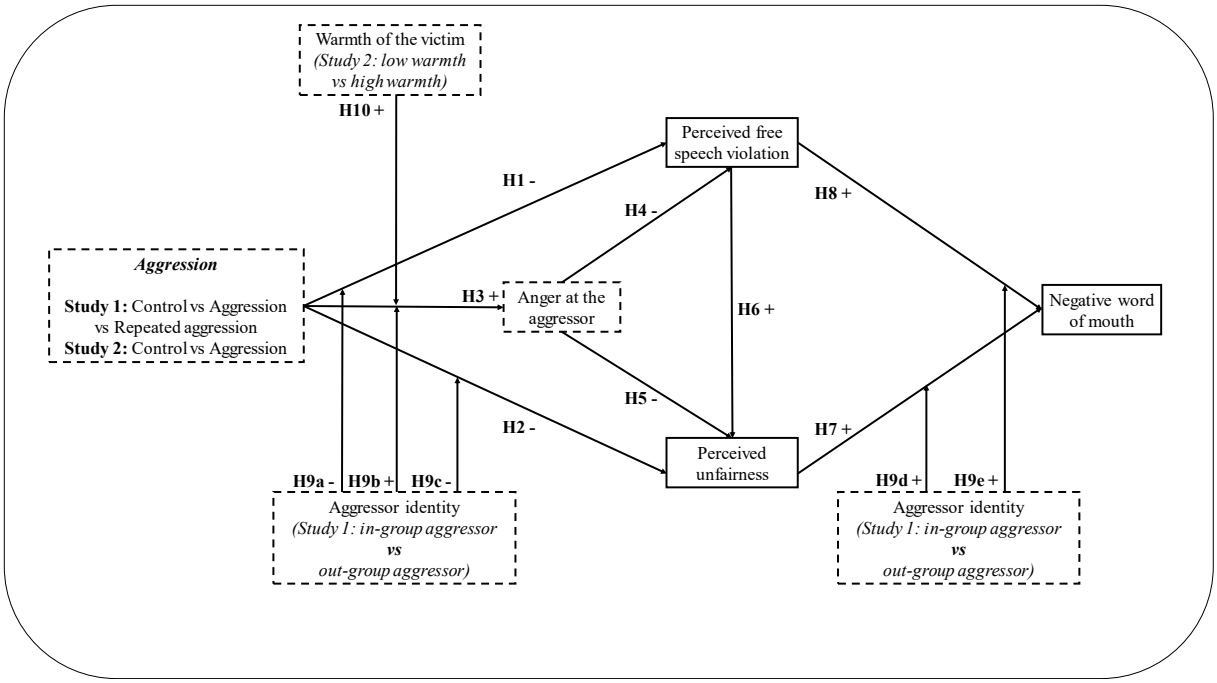
Wang, H. C., Doong, H. S., & Foxall, G. R. (2010). Consumers' intentions to remain loyal to online reputation systems. *Psychology & Marketing*, *27*(9), 887-897.

Wells, G. (2019, February 11). Writer sues twitter over ban for criticizing transgender people. *The Wall Street Journal*. Retrieved from https://www.wsj.com/articles/writer-sues-twitter-over-ban-for-mocking-transgender-people-11549946725.

White M. H. II, & Crandall, C. S. (2017). Freedom of racist speech: Ego and expressive threats. *Journal of Personality and Social Psychology*, *113*(3), 413-429.

Xie, C., & Bagozzi, R. P. (2019). Consumer responses to corporate social irresponsibility: The role of moral emotions, evaluations, and social cognitions. *Psychology & Marketing*, *36*(6), 565-586.

Yuksel, U., Thai, N. T., & Lee, M. S. (2020). Boycott them! No, boycott this! Do choice overload and small-agent rationalization inhibit the signing of anti-consumption petitions? *Psychology & Marketing*, *37*(2), 340-354.



Boxes with solid lines relate to the evaluation of the ban, boxes with dashed lines relate to the evaluation of the aggression.

Figure 1: Conceptual model

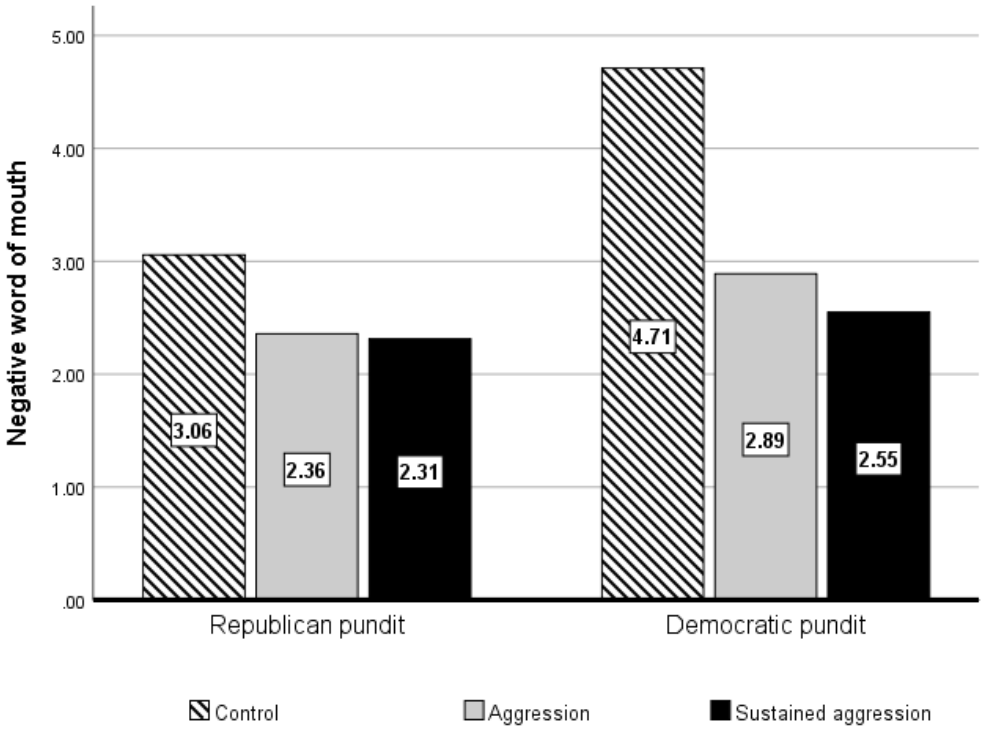


Figure 2: Interaction of aggression and aggressor identity on NWOM against the ban

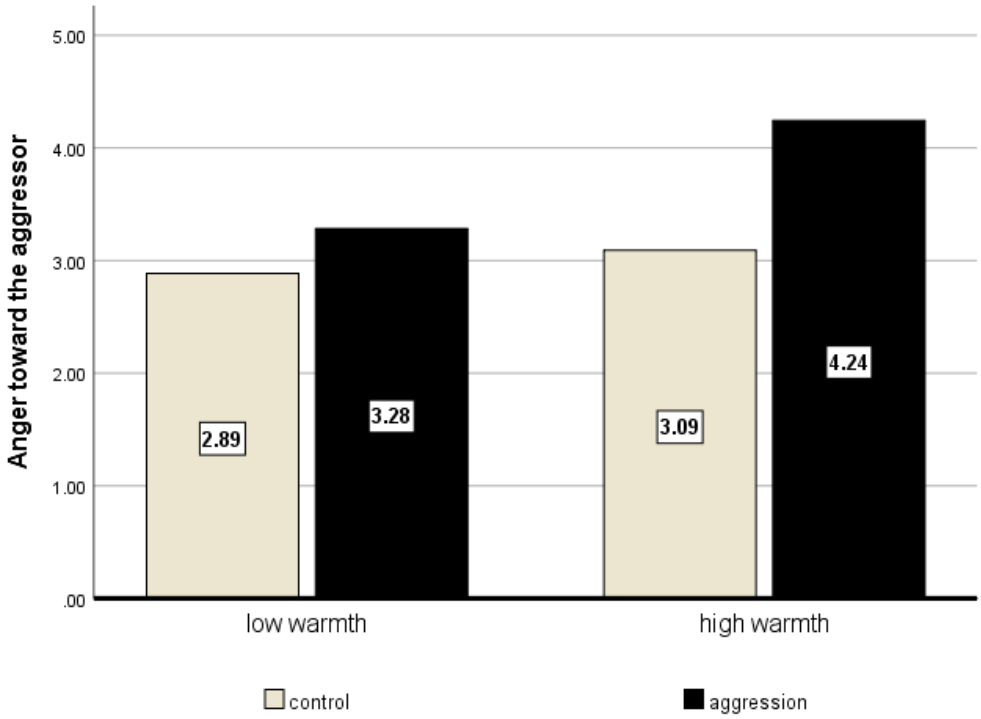


Figure 3: Interaction of aggression and victim's warmth on anger toward the aggressor

Table 1: Moderated-mediation model (Study 1 and Study 2)

Study	Parameters estimated	β	95% CI Lower	95% CI Upper
Study 1	Aggression → Anger toward the aggressor	.27*	.04	.49
	Aggression → Free speech violation	-.51**	-.74	-.28
	Anger toward the aggressor → Free speech violation	-.17**	-.31	-.03
	Aggression → Perceived unfairness of the ban	-.26**	-.42	-.11
	Anger toward the aggressor → Perceived unfairness of the ban	-.04	-.13	.05
	Free speech violation → Perceived unfairness of the ban	.76**	.67	.85
	Aggression → NWOM	-.26*	-.46	-.07
	Free speech violation → NWOM	-.01	-.18	.17
	Perceived unfairness of the ban → NWOM	.50**	.32	.68
	Aggressor's identity → NWOM	-.36	-.91	.18
	Free speech violation X Aggressor's identity → NWOM	-.18*	-.36	-.001
	Perceived unfairness of the ban X Aggressor's identity → NWOM	.37**	.19	.54
	R^2 = .50; F (8, 187) = 23.53, p < .001			
	Repeated aggression → Anger toward the aggressor	.40**	.17	.63
	Repeated aggression → Free speech violation	-.70**	-.93	-.47
	Anger toward the aggressor → Free speech violation	-.20*	-.34	-.07
	Repeated aggression → Perceived unfairness of the ban	-.31**	-.48	-.14
	Anger toward the aggressor → Perceived unfairness of the ban	-.12*	-.22	-.02
	Free speech violation → Perceived unfairness of the ban	.69**	.59	.79
	Repeated aggression → NWOM	-.23*	-.43	-.02
	Free speech violation → NWOM	-.02	-.18	.14
	Perceived unfairness of the ban → NWOM	.54**	.38	.70
	Aggressor's identity → NWOM	-.10	-.60	.40
	Free speech violation X Aggressor's identity → NWOM	-.19**	-.35	-.03
	Perceived unfairness of the ban X Aggressor's identity → NWOM	.31**	.15	.47
	R^2 = .51; F (8, 188) = 24.62, p < .001			
Study 2	Aggression → Anger toward the aggressor	.39**	.22	.55
	Warmth of the victim → Anger toward the aggressor	.29**	.13	.46
	Aggression X Warmth of the victim → Anger toward the aggressor	.19*	.02	.35
	R^2 = .10; F (6, 290) = 5.17, p < .001			
	Aggression → Free speech violation	-.26*	-.45	-.08
	Anger toward the aggressor → Free speech violation	-.16*	-.28	-.04
	R^2 = .09; F (4, 291) = 7.9063 p < .001			
	Aggression → Perceived unfairness of the ban	-.22*	-.36	-.08
	Anger toward the aggressor → Perceived unfairness of the ban	-.11*	-.19	-.02
	Free speech violation → Perceived unfairness of the ban	.61**	.52	.69
	R^2 = .48; F (7, 289) = 39.60, p < .001			
	Aggression → NWOM	-.09	-.25	.08
	Free speech violation → NWOM	.15*	.02	.28
	Perceived unfairness → NWOM	.53**	.40	.66
	R^2 = .39; F (5, 291) = 37.32, p < .001			

β represents unstandardized path coefficients. *p < .05; **p < .01. None of the covariates has a statistically significant effect on the variables examined. The aggression (repeated aggression) variable was coded as -1 control and +1 aggression. Warmth was coded as -1 in case of low warmth and +1 in case of high warmth. The average of all the items is used for the analysis. NWOM=negative word of mouth

Table 2: Conditional indirect effect analysis (Study 1)

	Hypothesized indirect effect	**Aggressor's identity**	**Coefficient**	**95% CI**
Independent variable: Control vs Aggression	Aggression → Free speech violation → NWOM	IG	.09	-.03 to .23
	Aggression → Free speech violation → NWOM	OU	-.09	-.24 to .04
	Aggression → Perceived unfairness → NWOM	**IG**	**-.23**	**-.41 to -.08**
	Aggression → Perceived unfairness → NWOM	OU	-.03	-.13 to .05
	Aggression → Anger → Free speech violation → NWOM	IG	.01	-.00 to .03
	Aggression → Anger → Free speech violation → NWOM	OU	.01	-.03 to .00
	Aggression → Anger → Perceived unfairness → NWOM	IG	-.01	-.04 to .02
	Aggression → Anger → Perceived unfairness → NWOM	OU	-.00	-.01 to .00
	Aggression → Free speech violation → Perceived unfairness → NWOM	**IG**	**-.34**	**-.53 to -.17**
	Aggression → Free speech violation → Perceived unfairness → NWOM	OU	-.05	-.18 to .07
	Aggression → Anger → Free speech violation → Perceived unfairness → NWOM	IG	-.03	-.08 to .00
	Aggression → Anger → Free speech violation → Perceived unfairness → NWOM	OU	-.00	-.02 to .01
Independent variable: Control vs Repeated aggression	Repeated aggression → Free speech violation → NWOM	IG	.15	-.01 to .29
	Repeated aggression → Free speech violation → NWOM	OU	-.12	-.29 to .03
	Repeated aggression → Perceived unfairness → NWOM	**IG**	**-.26**	**-.45 to -.09**
	Repeated aggression → Perceived unfairness → NWOM	OU	-.07	-.16 to .01
	Repeated aggression → Anger → Free speech violation → NWOM	IG	.02	-.00 to .05
	Repeated aggression → Anger → Free speech violation → NWOM	OU	-.01	-.04 to .00
	Repeated aggression → Anger → Perceived unfairness → NWOM	IG	-.04	-.10 to .00
	Repeated aggression → Anger → Perceived unfairness → NWOM	OU	-.01	-.04 to .00
	Repeated aggression → Free speech violation → Perceived unfairness → NWOM	**IG**	**-.41**	**-.60 to -.25**
	Repeated aggression → Free speech violation → Perceived unfairness → NWOM	OU	-.11	-.24 to .01
	Repeated aggression → Anger → Free speech violation → Perceive unfairness → NWOM	**IG**	**-.05**	**-.10 to -.01**
	Repeated aggression → Anger → Free speech violation→ Perceive unfairness → NWOM	OU	-.01	-.04 to .00

IG= In-group; OG=out-group; NWOM=negative word of mouth. Statistically significant indirect effects are highlighted in bold.

Table 3: Conditional indirect effect analysis (Study 2)

Hypothesized indirect effect	**Warmth of the victim**	**Coefficient**	**95% CI**
Aggression → Anger → Free speech violation → NWOM	LW	-.00	-.02 to .01
Aggression → Anger → Free speech violation → NWOM	HW	-.01	-.04 to .01
Aggression → Anger → Perceived unfairness → NWOM	LW	-.01	-.03 to .01
Aggression → Anger → Perceived unfairness → NWOM	**HW**	**-.03**	**-.07 to -.01**
Aggression → Anger → Free speech violation → Perceived unfairness → NWOM	LW	-.01	-.03 to .01
Aggression → Anger → Free speech violation → Perceived unfairness → NWOM	**HW**	**-.03**	**-.07 to -.01**

HW=high warmth; LW=low warmth; NWOM=negative word of mouth. Statistically significant indirect effects are highlighted in bold.

Appendix A: Scenarios used in Study 1

Control	
Democratic pundit - aggression	
Democratic pundit - repeated aggression	

Appendix B: Measures

Constructs	**Study 1**	**Study 2**
Free speech violation (1= Strongly disagree; 7= Strongly agree) **Study 1:** CR= .96, AVE= .90; **Study 2:** CR= .96, AVE= .90		
John M. Kenyon has a right to free speech so his tweet should not be banned	.96	.96
Twitter should respect free speech and not ban John M. Kenyon	.96	.96
Banning from Twitter users like John M. Kenyon is a violation of free speech that cannot be condoned	.93	.92
Perceived unfairness of the ban (1= Strongly disagree; 7= Strongly agree) **Study 1:** CR= .96, AVE= .89; **Study 2:** CR= .96, AVE= .88 Twitter's decision to ban John M. Kenyon is …		
Unfair	.95	.96
Unjust	.96	.88
Dishonest	.91	.96
Negative word of mouth against the ban (1= Strongly disagree; 7= Strongly agree) **Study 1:** CR= .96, AVE= .90; **Study 2:** CR= .96, AVE= .90 Given their decision to ban John M. Kenyon …		
I would be likely to complain about Twitter to other people	.91	.92
I would be likely to bad mouth Twitter to other people	.97	.97
I would be likely to say negative things about Twitter to people I know	.96	.96
Anger toward the aggressor (1= Strongly disagree; 7= Strongly agree) **Study 1:** CR= .96, AVE= .88; **Study 2:** CR= .97, AVE= .92 John M. Kenyon makes me feel:		
Angry	.95	.97
Mad	.96	.97
Indignant	.90	.93

Appendix C: Scenarios used in Study 2

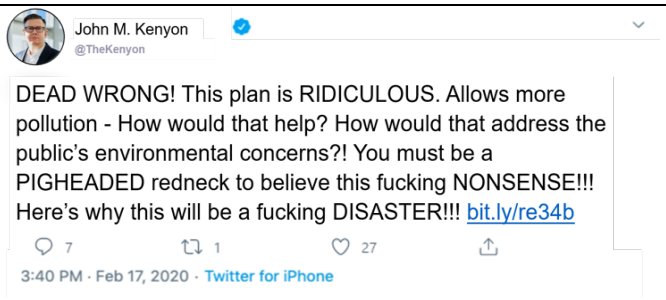
Control	
Aggression	
Low-warmth victim	Richard Matkey is Professor of Environmental Science and also acts as paid consultant to the boards of several gas and oil corporations and to their lobbies, providing scientific advice on a number of environmental issues. 
High-warmth victim	Richard Matkey is Professor of Environmental Science and also acts as volunteer for several environmental protection charities providing free scientific advice on a number of environmental issues. 

Figure legends

Figure 1: Conceptual model

Legend text: Boxes with solid lines relate to the evaluation of the ban, boxes with dashed lines relate to the evaluation of the aggression.

Figure 2: Interaction of aggression and aggressor identity on NWOM against the ban

Legend text: None

Figure 3: Interaction of aggression and victim's warmth on anger toward the aggressor

Legend text: None