April 2020

The Threat of Virality: Digital Outrage Combats the Spread of Opposing Ideas

Curtis Puryear
University of South Florida

The Threat of Virality: Digital Outrage Combats the Spread of Opposing Ideas

by

Curtis Puryear

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Psychology
College of Arts and Sciences
University of South Florida

Major Professor: Joseph Vandello, Ph.D.
Jennifer Bosson, Ph.D.
Jamie Goldenberg, Ph.D.
Sandra Schneider, Ph.D.
Michael Coovert, Ph.D.

Date of Approval:
March 24, 2020

Keywords: Morality, outrage, computer mediated communication, social media, political civility

ACKNOWLEDGMENTS

This dissertation was made possible by the contributions of many. I am grateful to my advisor, Dr. Joseph Vandello, for providing years of guidance and mentorship and to Dr. Jennifer Bosson for the years of helpful comments in lab meetings and brown bags. Thank you to my committee members, Drs. Jamie Goldenberg, Sandra Schneider, and Michael Coovert, who provided feedback to the ideas in this dissertation since they first took form in my major area paper. I am grateful to my master's adviser, Dr. Stephen Reysen, for introducing me to the world of psychological research and to Dr. Iva Katzarska-Miller, for connecting me with Stephen and setting me on my path to a career in research. Lastly, I would like to thank my parents, Donna and Carlton Puryear, without whose love and support I would not have survived the ups and downs of graduate school. The occasional checks in the mail helped too.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The prevailing stage for conversations about politics and morality has shifted from private and face-to-face to public and digital. Moreover, the digital landscape itself changed considerably in the past decade. The era of static webpages has been replaced by dynamic social networks where ideas and reactions to events spread rapidly. With every comment we, or a political adversary makes, numbers quantifying social approval tick up or down. Instead of holding digitized versions of one-on-one conversations, we argue in front of audiences who throw digital "points" at and accelerate the spread of the winning side's ideas. I argue this subjectively raises the stakes of moral and political discussions online, causing us to forego civility to combat the spread of ideas we oppose. Two experiments and one study of real-world interactions on Twitter test whether outrage and negative moral emotional language are triggered not only by the outrage inducing content on social media, but by their potential to spread and gain influence—to go *viral*. Furthermore, I test whether people use outrage strategically when trying to coordinate others against a target. Study 1 showed participants (N = 240) several animations of Tweets going viral (or not) in their first 12 hours. As predicted, outrage inducing content triggered greater subjective outrage and the desire to act when it went viral. Study 2 replicates this relationship in real world interactions between conservatives and liberals on Twitter (N = 22,092 tweet-reply pairs). In cross-ideological replies (e.g., liberals replying to conservatives), highly viral tweets attracted replies with twice the number of anger and negative-moral emotional words than non-viral tweets on average. No such relationship was observed in

homogeneous replies (e.g., liberals replying to liberals). Lastly, Study 3 explicitly instructed participants (N = 150) to either write replies to coordinate others against (i.e., downvote) another commenter or write replies they thought would cause others to reward (i.e., upvote) them personally. As predicted, explicit goals to coordinate audiences against a target triggered substantially more outrage expressions than attempts to gain personal rewards—even in the absence of changes in subjective outrage. Thus the viral spread of opposing ideas triggers outrage, which we use strategically to counter the threat of virality. In sum, talking about morality and politics with people who do not see the world as we do is already incredibly difficult. The present results suggest that "keeping score" of who is winning further impedes our chances at understanding one another.

INTRODUCTION

On a 2018 episode of his podcast, Ezra Klein, editor in chief for *Vox*, described a shift in the tone of online discussions. In the early days of the online "blog-sphere," he did not live in fear of scrolling down to the comments section below his posts. Disagreements felt sincere and non-threatening. As social media shifted from an era of static webpages to one of dynamic social networks filled with public markers of social approval, he felt a new kind of anxiety. Uncalculated miss-steps suddenly provoked mobs threatening attacks against his colleagues' and friends' reputations. People did not seem to argue to learn from one another anymore. They wanted to bring each other down, to see people barred from the conversation, even lose their jobs. Anecdotes like Klein's are easy to come by. People across the political spectrum describe growing concern over social media "pile-ons" in the overwhelming numbers made possible by social media. As the digital landscape shifted from static web pages to the modern era of dynamic social networks, it gained new social information—numbers quantifying social approval. Whether it be in the form of "likes," "retweets," or "upvotes," every major social media site broadcasts a score for the reception and impact of everything we post. Later in his podcast, Klein places the blame of rampant digital outrage squarely on the shoulders of these public markers of social reward. Coming to an understanding with people who do not share our worldview is already a difficult task, one that becomes even more difficult when it takes place in front of an audience throwing digital points at whichever side they think is winning. Social media takes conversations about morality that already contain aversive, opposing worldviews,

and embeds them in information that those views are spreading and gaining favor. In other words, social media raises the stakes of political conversation, causing us to resort to the nastiest tools in our repertoire to combat the worldviews we oppose.

I argue that social media encourages outrage in the service of *social coordination*. We condemn others to make a social impact, to change minds, and to rally people to our side of a dispute. We make moral appeals to resolve disputes over important decisions, like how to properly distribute resources or how to treat people who deviate from cultural norms about sexuality. Moral outrage and condemnation are two of our most powerful tools for shifting public opinion. The threat of opposing views going viral and spreading rapidly online may encourage us to use these tools more frequently in ideologically cross-cutting conversations.

Recently, others have argued that digital outrage is largely motivated by "*virtue signaling*," or the desire to appear morally good to others (Jordan & Rand, 2019). Introducing public markers of social approval into our interactions likely has multiple effects upon how we talk to one another. The *signaling* and *coordinative* functions of condemnation likely combine to account for the pervasiveness of outrage and shaming on social media. Thus, I argue these are complementary, rather than competing perspectives. However, while a variety of work attributes digital outrage to virtue signaling motivations (Grubbs, Warmke, Tosi, James, & Campbell, 2019; Johnen, Jungblut, & Ziegele, 2018; Jordan & Rand, 2019), little work has examined its coordinative function. I review both the signaling and coordinative motivations for digital outrage and explore how the modern era of social media facilitates both. Then, I conduct three studies providing an initial demonstration of the *coordinative* roots of digital outrage. More specifically, I demonstrate that 1) seeing opposing coalitions form online prompts outrage in both an online experiment and in real world conversations on Twitter and 2) people use outrage

strategically when they have an explicit goal to coordinate others against a target. Combined these studies provide initial evidence that digital outrage is triggered by the virality of our political adversaries and employed strategically to combat their spread.

OUTRAGE AS VIRTUE SIGNALING

One of the key insights of early research studying online interactions is the control they provide over self-presentation (Wallace, 1999). Whether it be through Snapchat filters, the pictures we post on Instagram, or in the information we include in our Facebook bios, social media lets us choose what aspects of ourselves we show the world. Digital networks also let us control the information that invades our news feeds. We choose who to follow, friend, block and talk to on sites like Facebook and Twitter. This enhanced control over who we interact with can be especially liberating and make us feel more comfortable sharing our true selves with one another (McKenna, Green, & Gleason, 2002). Moral traits lie at the center of our true selves (Newman, Bloom, & Knobe, 2014), and the Internet provides ubiquitous opportunities to talk about our values. People witness more immoral acts online than in person or via traditional media (Crockett, 2017), roughly half of Americans report being civically active on social media, and 37% say social media is an important venue for expressing their political convictions (PEW, 2018). The sudden explosion of opportunities to communicate our moral traits to like-minded others and reap social rewards offers one explanation for the rise of outrage culture.

Partner choice models of human interaction help explain why our moral reputations are so important to us (Baumard, Andre, & Sperber, 2013). Researchers often examine cooperation using economic games, or lab studies in which groups of people make decisions about how to distribute resources. One version of the Prisoner's Dilemma, for example, assigns participants to interact with a single interaction partner in which they can choose to cooperate or act selfishly

over repeated trials. An effective tactic for encouraging cooperation in is to engage in a *tit-for-tat* strategy (Axelrod & Hamilton, 1981), cooperating when one's partner does so and punishing selfish behavior by returning the favor. Tit-for-tat strategies reflect a narrow type of morality, one centered on reciprocity. But Baumard and colleagues (2013) argue this paradigm cannot explain the emergence of human's moral sense in its entirely. Partner choice models (Barclay, 2016; Noe & Hammerstein, 1993) posit that taking an "eye for an eye" is often not our only recourse when faced with a selfish partner. In many, if not most exchanges, we also have the option of choosing another interaction partner with a better reputation for cooperation. Under models allowing for partner choice, one's reputation suddenly acquires a great deal of value (Fu, Hauert, Nowak, & Want, 2008). Reaping the benefits of cooperation depends upon successfully signaling you are a dependable exchange partner who will share costs and benefits equally.

Both historical and experimental evidence demonstrate that partner choice encourages cooperation. From traders in medieval Europe (McAdams, 1997) to Jewelers in New York (Bernstein, 1992), partners throughout history have made deals even in the absence of judicial oversight. The irreparable costs of exclusion from cooperation motivate people to deal fairly. When other partner options are present, developing a reputation for unfair transactions leads potential partners to choose others when exchanging good and services. Moreover, experimental evidence consistently finds that people choose exchange partners based on their reputation for cooperation (Barclay, 2006; 2013; Barclay & Willer, 2006; Rockenbach & Milinski, 2011; Santos, Rankin, & Wedekind, 2013; Sylwester & Roberts, 2010). People can effectively infer their partners' penchant for cooperation by tuning into details like how spontaneously they choose to behave pro-socially (Verplaese, Vanneste, & Braeckman, 2007). People also predict better than chance whether a future partner will cooperate in a one-shot prisoner's dilemmas if

they have an unrelated conversation beforehand (Brosig, 2002). Because our potential as cooperation partners is under constant surveillance, communicating our reputation as a fair interaction partner requires vigilance. So much vigilance in fact, that Baumard and colleagues (2013) argue genuine concerns for fairness emerged to motivate behavior signaling our suitability as exchange partners.

Consistent with the benefits of reputation signaling, we present ourselves in ways that accrue benefits, recognition, and favorable views (Baumeister, 1982; Leary & Kowalsky, 1990; Schlenker & Weigold, 1992). Moral traits hold special status in self-presentation. Feeling judged as immoral carries greater weight than being judged as incompetent (Leach, Ellemers, & Barreto, 2007), people are more likely to exaggerate the morality of their own behavior than their intelligence (Allison, Messick, and Goethals, 1989), they are willing submerge their hands in worms and icy water in order to avoid damage to their reputations (Vonasch, Reynolds, Winegard, & Baumeister, 2017), and changes in moral traits have a significantly larger impact on self-perceived identity than changes in personality traits, memories, preferences, basic cognitive capacities, perceptual abilities, and physical features (Strohminger & Nichols, 2014; 2015). Groups respond defensively to moral identity threats (Sullivan, Landau, Branscombe, & Rothschild, 2012), they compete more fiercely for moral than material status (Leach, et al., 2007), and they compete for victim status in order to gain the moral high ground (Young & Sullivan, 2016). People care more deeply about how they are perceived in moral terms than perhaps any other qualities.

Humans take advantage of what signals are available to communicate their moral reputations. Given the opportunity to make moral judgments in front of others, demonstrating a committing to moral duty (i.e., deontological judgments) reliably signals trustworthiness to

audiences (Everett, Faber, Savulescu, & Crockett, 2018; Everett, Pizarro, & Crockett, 2016; Rom & Conway, 2018; Uhlmann, Zhu, & Tannenbaum, 2013). Everyday life is not filled with moral dilemmas for us to solve publicly; however, prosocial behavior, such as sharing, is a powerful signal for reputation that brings rewards (Jordan & Rand, 2017; 2019). In the absence of opportunities to behave prosocially we often take advantage of a signal that stands in stark contrast to helping behavior: condemnation. In economic games, when participants do not have the option to share resources as a signal of trustworthiness, participants' look to each other's tendency to punish selfish players when deciding with whom to cooperate (Barclay, 2006; Jordan et al., 2016; Nelissen, 2008). Moreover, choosing cooperation partners based on their punishment history actually leads to better outcomes for cooperation (Jordan et al., 2016). Third-party punishment provides a viable option to signal trustworthiness to potential cooperative partners, and people increase their punishment of transgressors in front of audiences (Kurzban, DeScioli, & O'Brien, 2007).

In short, we care deeply about whether others view us as good and fair, we go to great lengths to protect our reputations, and doing so brings a host of social rewards. Theories drawing from partner choice models argue the centrality of our moral identities stems from their ability to boost our chances of being chosen as cooperation partners. We employ several strategies for signaling our reputations to others, one of which is punishing moral transgressors in front of audiences. While multiple options for signaling moral traits exist, the Internet may be especially effective at increasing the viability of condemnation as a signaling strategy.

Virtue Signaling in Digital Space

Communications researchers, social psychologists and organizational psychologists have all written on the defining characteristics of computer mediated and digital communication. This

literature stretches as far back as theories of social presence in communication science over four decades ago (Short, Williams, & Christie, 1976). Social presence theory and media richness theory (Daft & Lengel, 1986) focus on deficiencies in nonverbal, paraverbal, and other social context clues in computer mediated communication. More recently, social psychologists Katelyn McKenna and John Bargh (2002) highlighted four differences in communication over the Internet compared with face-to-face interactions: social anonymity, the irrelevance of physical distance for interaction partners, the unimportance of physical appearance and visual cues for relationship formation, and greater control over the time and pacing of social interaction. Finally, organizational psychologists emphasize the richness of information transmitted and the synchronicity of communication as key dimensions (Kirkman & Mathieu, 2005).

But computer mediated environments have changed vastly since researchers began investigating the effects of communicating via email in organizations. The Internet has transformed from an era characterized by static web pages in the early 2000s to the constantly shifting and reacting nature of social media. The characteristics of modern digital environments may facilitate the signaling function of outrage expression. First, web users have considerable control over what groups they choose to engage with online, making it easier to selectively express outrage among groups who will respond positively. Second, the Internet provides near unlimited access to morally relevant discussions that reward outrage expression. Social media provides concrete indicators of social approval in the form of "likes" and "upvotes" (i.e., buttons users press to show their approval to others) that give concrete, quantified feedback for moral reputation unlike anything in face-to-face reactions (in which approval must either being inferred from others behavior or is simply not available because the vast audiences of digital networks are impossible).

Opportunities for Expression and Reward

In a reanalysis of data tracking everyday moral experiences (Hofmann, Wisneski, Brandt, & Skitka, 2014), Crockett (2017) finds that people are more likely to learn about immoral acts online than in person or via traditional media. Moreover, immoral acts encountered in digital media tend to be more extreme and outrage inducing than the more mundane transgression (e.g., being cut off in traffic, seeing someone jaywalk, etc.) encountered in face-to-face contexts. Prior to the advent of digital media, information about people's moral character spread through our local social networks via gossip to inform judgments about trustworthiness and cooperation with others in local communities. Much of the information we encounter about the moral character of companies, politicians, celebrities, or other public figures comes from news organizations seeking to gain traffic ad revenue from users. This provides a financial incentive for the creation of especially outrageous "click bait" to attract web users to sites that depend upon "clicks" for ad revenue. The steady, repeated exposure to especially outrage-inducing content in digital media, which one would rarely otherwise encounter in person, offers numerous opportunities to condemn others.

Comments on social media are also met with concrete social feedback. Social approval activates the reward centers of the brain (Meshi, Morawetz, & Heekeren, 2013; Sherman, Payton, Hernandex, Greenfield, & Dapretto, 2016), facilitates learning (Ruff & Fehr, 2014) and boosts self-esteem (Burrow & Rainone, 2017). Beyond the immediate rewards of digital feedback, Crockett (2017) argues that people deliver "likes" and "favorites" in patterns resembling variable interval reinforcement schedules which are especially effective at forming habits (Dickinson, Nicholas, & Adams, 1983). Thus social media provide ubiquitous opportunities for reputation reinforcement, which may produce widespread expressions of

outrage out of habits engrained by unpredictable patterns of social rewards. In short, social media amplify both the opportunities and rewards for signaling our moral traits through condemnation.

Audience Filtering Online

The control social media grants its users over social interaction may make condemnation a more viable strategy for signaling virtues. Punishing free riders or moral norm violators often provokes retaliation (Herrmann, Thoni, & Gachter, 2008). Individuals who punish others for unfairly allocating resources in economic games often receive retaliatory punishments, a phenomenon labelled "antisocial punishment" (Rand & Nowak, 2011). Moreover, publicly condemning a divisive position risks alienating those with dissimilar attitudes. Even people who are most supportive of diversity broadly withhold their tolerance for moral diversity (Haidt, Rosenberg, & Hom, 2003), and people attribute bad character to targets who disagree with their moral judgments about disgust inducing transgressions (Katzir, 2017). Risks of retaliation from detractors may be especially large in today's political climate as views towards ideological opponents have grown increasingly negative in recent decades (PEW, 2016). While public condemnation holds potential benefits (Barclay & Kiyonari, 2014; Jordan & Rand, 2017; Jordan, et al., 2016; Raihani & Bshary, 2015; Santos et al., 2013) the costs of public punishment limits its prevalence in traditional social interactions (Guala, 2012). However, digital environments may sidestep a number of these risks while preserving the utility of punishment for signaling moral reputation. Social media grants users increased control over the audiences of their condemnation, allowing users to segregate themselves into ideological bubbles of like-minded others. Furthermore, fears of uncomfortable social interactions or even physical retaliation are less tenable for interactions mediated by a computer screen. Thus the cost-benefit ratio for condemnation may become more favorable in digital contexts.

The Internet allows us to selectively expose ourselves to information that agrees with our worldview. People consistently prefer associating with similar others. We embed ourselves in social networks comprising homogenous sociodemographic, behavior, and intrapersonal characteristics (McPherson, Smith-Lovin, & Cook, 2001) and migrate so our neighbors share our political views (Motyl, Iyer, Oishi, Trawalter, & Nosek, 2014). Consumers gravitate to news from sources that share their ideology (Iyengar & Hahn, 2009; Munson & Resnick, 2010), and news aggregators take advantage of this by recommending content that aligns with our ideology (Pariser, 2011). Some commentators, such as Cass Sunstein, have predicted digital communities to form echo chambers of like-minded individuals reinforcing one another's views and fostering polarization (2018). While evidence that digital echo chambers are undermining democracy or fostering extreme opinions is mixed (Tucker et a., 2018), digital environments do appear to at least provide the tools for filtering out attitudinally dissimilar others.

Social media may feed our homophilous tendencies even further than traditional media. We naturally tend to follow and friend more like-minded others without deliberation (Aiello et al., 2012). Moreover, websites like Twitter give users direct control over the content appearing in their newsfeeds, allowing them to exclusively follow ideologically similar accounts if they choose. Facebook lets users mute disagreeable friends and make one's posts invisible to specific people. Other discussion forums like reddit.com feature sub-forums created specifically for people of a given ideology. For example, the wiki for r/conservative, Reddit's forum discussing conservative perspectives, describes itself in the following manner:

> "We are not *fair and balanced*. We don't pretend to be unbiased. We don't pretend to give all commenters equal time. This is *by conservatives and for conservatives*. We are here to

> discuss conservative topics from a distinctly conservative point of view. If you don't like that it's not an unbiased forum, go ask why /r/politics is a leftist totalitarian state. Leftists and moderates have never been welcomed here. If you wander in here and spout nonsense or insult us, don't be surprised when we ban you almost instantly."

While the most comprehensive analyses do not suggest mass ideological segregation online (Eady, Nagler, Guess, Zilinsky, & Tucker, 2019; Tucker et al., 2018), examples of web users taking advantage of social media's filtering tools is well-documented. Twitter users tend to form politically homogenous clusters (Himelboim, McCreery, & Smith, 2013), and political tweets are retweeted more frequently within (vs across) ideological groups (Brady, Wills, Jost, Tucker, & Van Bavel, 2017). Liberal Facebook users' friend networks comprise less than 20% conservative and more than 60% liberal, while conservatives' networks mirror this pattern almost perfectly (Bakshy, Messing, Adamic, 2015). Moreover, just as third parties are more likely to punish moral transgressors face-to-face when punishment is backed by consensus (Konishi, Oe, Shimizu, Tanaka, & Ohstudbo, 2017), so too are Facebook users selectively less willing to condemn others on social media if they believe their followers disagree with them (Hampton, Rainie, Dwyer, Shin, & Purcell, 2014). People who sense incongruency between their opinion and the national climate also report less willingness to post comments on moral issues (Gearhart & Zhang, 2014).

Examinations of participation in online firestorms (i.e., collective panics in response to perceived threats to cherished values) find users selectively comment to maximizes potential reputational payoff (Johnen et al., 2018). Given a high volume of outrage directed at a moral norm violation online, participants become less willing to write a comment of their own. The

authors argue that high volumes of previously expressed outrage undermine opportunities to stand out and gain social recognition, rendering outrage an ineffective signal of personal reputation. This also helps explain why online fire-storms are typically short lived (Pfeffer, Zorbach, & Carley, 2014). Field studies on Twitter support the negative relationship between pre-existing comments and willingness to comment oneself. As the number of pre-existing replies increases, previous commenters become less likely to comment again on a Twitter thread covering a political topic (Shugars & Beauchamp, 2019). In other words, if a user replies to a Tweet, leaves for some time, then returns to find the Twitter thread has received many comments, they are less likely to comment again than if a relatively small number of replies had been made. Again, one interpretation of these results is that users choose not to express outrage or engage in heated arguments when their comments are likely to be lost in the crowd.

Summary: Digital Outrage as a Reputation Signal

The signaling perspective offers one explanation of why outrage culture feels pervasive on social media. We care deeply about our moral reputations and public condemnation effectively signals our moral qualities. While traditional, face-to-face exchanges provide limited and risky opportunities for condemnation, digital networks provide an unlimited supply of transgressors to safely condemn in our networks of like-minded others. From this view, mobs expressing extreme degrees of outrage stem from their members' attempts to make their signals stand out from the crowd. We condemn others and express outrage to amplify the signal of our reputation. This does not imply that outrage is feigned or not genuinely felt (Jordan & Rand, 2019). Part of the reason we *feel* our moral convictions so intensely is because the motivation to signal and protect our reputations is adaptive (Baumard et al., 2013). To the extent that outrage culture is motivated by signaling, homophily may dominate social networks in order to maximize

the chance of outrage expression resulting in social rewards. While “echo chamber” does not accurately describe large segments of digital networks, considerable evidence demonstrates that people do frequently take advantage of the filtering capabilities of digital media.

OUTRAGE AS SOCIAL COORDINATION

The signaling perspective excels at explaining outrage in "echo chambers." If outrage culture takes root in the motivation to reap social rewards from publicly condemning our opponents, then surrounding ourselves with people who share our convictions will yield those rewards most consistently. But many interactions on social media are between ideological opponents (Eady et al., 2019; Lee, Choi, Kim, & Kim, 2014; Yardi & Boyd, 2010). Moreover, outrage also exists in completely anonymous platforms (e.g., Reddit) where it is impossible to signal personal reputation. Proponents of the signaling perspective argue that even in anonymous contexts, we still engage in costly punishment as a heuristic (Jordan & Rand, 2019). In other words, even in situations where condemnation cannot boost our reputations, we still try to signal them via condemnation as a general rule. Alternatively, the functions of outrage expression may extend beyond virtue signaling. Surely sometimes people argue with and condemn others because they are genuinely motivated to shape the moral rules that constrain and reward how people treat one another. Outrage is often our best tool for fighting for the principles we believe in, for undermining the reputations of our opponents, and rallying allies to our side in moral disagreements. Condemnation is not only a signal of personal virtue; it is how we dictate the sides people choose in conflict.

The Side-Taking Perspective on Morality

In the film *Black Panther*, the people of Wakanda, a fictional civilization in sub-Saharan Africa, enjoy a host of fantastic, futuristic technologies—from levitating chariots and

superpowered suits to bracelets capable of healing gunshot wounds. At the same time, Wakanda employs a much older method for settling competing claims to the throne: trial by combat. In the real world, ancient judicial systems used trials by ordeal in which the accused were subjected to some painful, usually dangerous experience, and escaping unscathed was taken as proof of innocence. The outcome of the trial makes it easy for groups in disagreement to choose a side. Throwing an accused witch into a lake to see if they float as a test of guilt creates a visible signal for social coordination. Similarly, pitting two would-be kings against one another in a fight to the death makes it easy to choose a ruler afterwards. But fights to the death have limitations, like allowing a murderous villain to ascend to the throne by out-dueling a more caring and wiser opponent. If you have watched the pivotal scene in *Black Panther* where this happens, you might have heard a voice in your head screaming, "Forget the trial by combat! Don't give the villain the crown because he's a horrible, horrible person!" Morality is an incredibly useful tool for settling disputes. Choosing a side not because they are powerful or because they are our friends but *because they are good* often results in superior outcomes. Of course the people of Wakanda do use their moral sense to coordinate against the villain by the end of the film, but not soon enough to avoid the costs of side-taking strategies void of moral input.

DeScioli and Kurzban's (2013) side-taking model of morality argues condemnation is one of our best tools for navigating disputes over resources, rulers, and all the other potentially costly disagreements humans encounter. Moral condemnation has clear advantages over other strategies for social coordination. In the absence of moral appeals, humans typically choose sides based on whichever disputant has the most power or based on pre-existing alliances with one of the disputants. Both these strategies bring unique limitations. The former lowers the cost of conflicts but ultimately leads to despotism with the same individuals consistently reaping the

greatest rewards and maintaining power. The latter often leads to costly conflicts between sides entrenched in alliances. DeScioli and Kurzban (2013) propose that moral appeals allow for more dynamic coordination. Granting one side the moral high ground creates a signal visible to onlookers that side-steps the traps of despotism and justifies forsaking pre-existing alliances.

In support of their condemnation centered model, DeScioli and Kurzban point to observations that moral judgments are frequently not based on promoting group welfare, that they track rule violations rather than benefits (Mikhail, 2007), that moral principles themselves are frequently damaging to groups (Ryan, 2014), and that behavior is far more motivated by appearing moral than by actually following moral principles (DeScioli & Kurzban, 2009). Condemnation is a weapon capable of destroying opponents' reputations, depriving them of friends, and recruiting their former allies to one's side. Moral disgust motivates people to avoid targets of blame (Curtis & Biran, 2001; Hutcherson & Gross, 2011; Molho, Tybur, Guler, Balliet, & Hofmann, 2017; Tybur, Lieberman, & Griskevicius, 2009), moral condemnation puts one at risks of exclusion and exploitation from communities (Opotow, 1990), moralization drives groups to act on their sides' cause, increasing political action and the acceptability of violent means to achieve morally justified ends (Skitka, 2010; Skitka, Bauman, & Sargis, 2005), and lacking social support predicts an increased tendency to moralize or invoke condemnation in an attempt to gain social support (Peterson, 2013). Invoking morality both has the power to punish opponents and to rally allies in disputes.

Fighting Over Moral Rules

For moral judgments to coordinate behavior effectively, we need to agree upon the moral rules that will tell us which side is right. Descioli and Kurzban refer to these arguments over which rules will coordinate behavior as "moral meta-fights." Both across time and culture, fights

over moral rules have produced diverse sets of principles to guide behavior. Views towards corporal punishment, slavery, civil rights, human sexuality, and animal rights have shifted considerably over human history (Pinker, 2010), and local moralities continue to show extraordinary cross-cultural heterogeneity (Tooby & Cosmides, 2010). This also allows moral rules to prohibit harmless or beneficial behaviors, like interest bearing loans or same-sex relationships—if the people who benefit from these rules can convince their communities to adopt them. But destructive moral norms are not bugs in the system. They demonstrate the flexibility of the moral domain and its susceptibility to being co-opted in the service of diverse coordination goals.

Moral rules cannot coordinate effectively amidst disagreement. This is partly why we find moral disagreements so aversive. We do not tolerate moral diversity (Haidt, et al., 2003), and we shun close neighbors who embrace moral relativism (Sarkissian, Park, Tien, Wright, & Knobe, 2011). Countries often contain groups who have reached different conclusions about the rules that will coordinate their behavior. In the United States the political right and left have reached consensus on competing rules for things like sexual relationships, how to fairly redistribute wealth, and our duties towards non-citizens. Again, people will often go to great lengths to simply avoid moral disagreement, even foregoing money (Frimer, Skitka, & Motyl, 2017). Other times this creates conflict, tugs-of-war over the rules that will dictate behavior. Moral outrage and condemnation are some of our most important tools in our fights over moral rules.

Why do we care about which behaviors our community rewards and punishes? Why bother arguing about morality? Why not just conform to whatever rules dominate our immediate surroundings? The outcomes of moral rules benefit and harm some more than others. In other

words, we care about which principles guide behavior because we often have skin in the game. For example, someone with a history of and predilection for sexual promiscuity would suffer if their culture suddenly adopted restrictive rules about casual sex. Likewise, the wealthy benefit from moral narratives describing taxes as government sanctioned theft. We appeal to morality constantly to coordinate people to our cause in disputes. For our rules to effectively coordinate behavior, they must hold some degree of consensus. So we argue, and we fight, and we condemn those who transgress against them.

Consistent with a motivation to fight for specific moral rules (rather than conform blindly), evidence from across the moral domain suggests our judgements are often tied to self-interest. People often engage in behavior they condemn in others when it benefits them (Batson & Thompson, 2001), and they judge people who offer them benefits more leniently than individuals who do not (Bocian & Wojciszke, 2014). People selectively endorse moral principles to support their pre-existing beliefs about racism (Uhlmann, Pizarro, Tannenbaum, & Ditto, 2009). They judge resource distributions that offer more generous payouts to their role in a collaborative task as more fair (DeScioli, Massenkoff, Shaw, Peterson, & Kurzban, 2014). Views on abortion and recreational drug use are both partly explained by individual differences in reproductive strategies (Kurzban, Dukes, & Weeden, 2010; Quintelier, Ishii, Weeden, Kurzban, & Braeckman, 2013; Weedon, 2003). And membership in higher status groups predicts endorsing ideologies that legitimize and maintain group-based hierarchies (Pratto, Sidanius, Stallworth, & Malle, 1994).

Of course all our moral convictions are not grounded in self-interest. We can fight for moral rules because they benefit others we care about or even because hours of sitting in our philosophical armchairs lead us to believe certain principles will create the best world to inhabit.

Rules with consequences for our self-interest are the most obvious examples of when we fight over moral judgments, but the motivation to push for a specific side of a moral dispute can stem from a variety of sources.

Side-Taking in Digital Space

When we talk about politics and morality face-to-face, we typically know each other's identities. We see all our opponents' idiosyncrasies that make them feel like individuals. In person arguments about politics often involve a few people talking in private. Condemning someone in this context risks alienating close friends who might be listening or provoking retaliatory attacks. Moreover, the target of our outrage feels like a real person, not an avatar representing the conservative or liberal agenda. But what if we put masks on both sides of a dispute, set them on a stage in front of thousands, and hang point totals over their heads tallying which side has won the most support from onlookers? This increasingly describes the context of political discourse. Virtual environments decrease the risks of condemnation, they transform unique individuals into homogenous, moral opponents or allies, and they keep a running a running "score" for who is winning the battle over the moral high ground. In face-to-face discussion, fighting over moral rules is costly, has limited reach, and is inhibited by our perceptions of one another's humanity. Digital space removes these obstacles, making condemnation a more viable coordination strategy.

Anonymity

Many digital environments detach users' online personas from their "real world" identities. Anonymity severs the impact of behavior upon personal reputation. This may disinhibit behavior online (Suler, 2004), allowing people to self-disclose personal information without feeling vulnerable (Bargh & McKenna, 2002) and to behave uncivilly without risking

reputational damage (Omernick & Sood, 2013). On the other hand, it may also render the signaling function of moral condemnation (Jordan & Rand, 2017) less effective. We cannot signal our moral character when our identities are hidden. Thus the impact of anonymity upon condemnation is tragic in a certain sense. It allows individuals to escape the risks of attacking other's character, but it also blocks the benefits of condemnation for our own reputation. However, anonymity poses no obstacles for outrage in the service of coordination. In fact just the opposite, anonymity grants us the freedom to use all the nastiest tools for manipulating moral consensus with none of the typical costs of doing so.

Calling others out is social risky. Punishing free riders or moral norm violators often provokes retaliation (Herrmann, et al., 2008). Punishing selfish behavior often generates retaliatory punishments, a phenomenon called "anti-social punishment" (Rand & Nowak, 2011). Public condemnation reveals our moral convictions, risking the possibility that our peers might hold conflicting beliefs. Violations of sacred values are typically met with intolerance and at times dangerous responses (Fiske & Rai, 2014). Anonymity minimizes these costs, while having no obvious effects on the efficacy of outrage as a coordination device.

The Salience of Sides

In the absence of identifying information, group identities often take over. Deprived of individuating cues, virtual interactions shift attention to others in terms of their similarity to prototypical group members (Lea, Spears, de Groot; 2001) and increase ingroup attraction and susceptibility to stereotyping and discrimination (Postmes, Spears, & Lea, 2002). When we have access to fewer social and biographical cues, we rely more upon our beliefs about the groups people belong to (Reicher, Spears, & Postmes, 1995). Thus as we enter the digital world, it often becomes easier to see each other according to group membership. For topics about politics and

morality, those group identities often signify disagreement about which rules should coordinate behavior. As we scroll through feeds of different Twitter handles and text arguing about politics, we may care less about who people are as individuals and more about "which side they are on."

Furthermore, perceiving others as homogenous members of an outgroup has nasty effects on intergroup relations. It increases discrimination (Vandeselaere, 1991), facilitates the use of aggression by casting outgroups as uniformly evil (Ostrom & Sedikides, 1992; Wilder, 1986) increases ingroup favoritism (Simon, 1992), and predicts seeing the outgroup as more threatening (Corneille, Yzerbyt, Rogier, & Buidin, 2001; Rothgerber, 1997). Socially deprived interactions online promote seeing each other in terms of the sides we take, which can facilitate using attacks and derogation against one another.

Consensus information

Prior to the Internet, the spread of moral reputations depended upon much slower, more localized mechanisms such as gossip (Piazza & Bering, 2008). Institutions devised methods to rapidly communicate the reputations of deviants or opposing groups through organized propaganda or public executions and trials (DeScioli & Kurzban, 2013). Now the U.S. news cycle is filled with stories of a foreign government, Russia, leveraging the consensus shaping power of social media to disrupt democracy. Farms with thousands of bots and fake accounts are used to spread misinformation and to shape what beliefs appear normal (Broniatowski et al., 2018; Badawy, Ferrara, & Lerman, 2018). Social media facilitate these efforts by design. Televised or printed propaganda have no "retweet" or "like" buttons. A comprehensive analysis of 41.7 million Twitter profiles in 2010 found that a single retweet alone leads to an average audience of 1,000, regardless of how many followers the original "tweeter" had (Kwak, Lee, Park, & Moon, 2010). Share and retweet buttons facilitate the rapid diffusion of information and

its evaluation, often called *virality* (Alhabash & McAlister, 2015). Even messages from sources with relatively few followers have the potential to go viral in the right time and place.

Social media, almost by definition, editorialize news. Every story is embedded in commentary from the poster and receives a score for social approval. Likes and the dreaded Twitter ratio (i.e., a high ratio of replies/favorites for a Tweet is taken to represent disapproval) provides a quantified signal of an audience's consensus. While likes and upvotes feel rewarding and tell us when we have successfully communicated our moral traits—hence their relevance to virtue signaling—they also tell us which side in a dispute currently holds consensus. They allow us to see if an opposing side is rapidly gaining social support versus eliciting backlash. If moral condemnation is about coordinating people to specific sides in disputes, then digital likes tell us which side is winning. In many ways arguing about politics face-to-face is like competing in a game without keeping track of each teams' points. Conversely, every argument we make on social media has the potential to suddenly go viral and wildly swing the score in our side's favor.

Evidence for Coordinative Outrage Online

In some ways social media resemble ideological echo chambers. Analyses of Twitter reveal that users are more likely to follow like-minded than dissimilar others (Halberstam & Knight, 2016; Hayat & Samuel-Azran, 2017; Himelboim, 2014) and are more likely to "retweet" messages from those who share their ideology (Brady et al., 2017; Himelboim, McCreery, & Smith, 2013). However, other examinations of digital networks reveal ubiquitous interactions between ideological opponents. Social media use is positively associated with exposure to politically diverse information (Bae, 2013) and an analysis of Facebook finds that more than 20% of users' friends hold opposing views (Bakshy et al., 2015). The most comprehensive analysis of "political bubbles" on Twitter finds that the ideological distributions of accounts

followed by extreme conservatives and liberals overlap by 51% (Eady et al, 2019). Social media is both homophilous and cross-cutting. This is consistent with multiple functions of online discussions about morality. Sometimes we surround ourselves with like-minded others to reap social rewards and boost our reputations with more certainty. Other times we argue and fight to undermine opponents and shift consensus behind our side's moral rules.

When researchers examine replies to original tweets (i.e., writing out a response to another user) rather than likes or retweets (i.e., actions only requiring single clicks that signal approval for or share others' posts), ideologically crosscutting interactions increase in prevalence. In other words, effortless sharing or liking reflects more homogeneity than deliberative replies (Liang, 2014). An analysis of Tweet-reply pairs following the shooting of late-term abortion doctor, George Tiller, produced substantial cross-ideological talk, with 396 out of 1,137 replies representing responses to opposing viewpoints (Yardi & Boyd, 2010), and like-minded Tweet-reply pairs constituted only 20% to 40% of total replies during the first 24 hours following the incident. Moreover, Twitter replies are more emotionally intense than original messages, but this pattern is almost entirely driven by course-correction rather than amplification. In other words, emotional escalation on Twitter is largely driven by negative responses to positive Tweets rather than like-minded response with increased emotional intensity (Goldenberg, Gross, & Garcia, 2018). Thus evidence for echo chambers versus crosscutting exposure likely depends upon the indicators of social media participation (e.g., written replies versus shares) researchers choose.

Analyses looking specifically at comment sections attached to online news articles find further evidence of cross-cutting political exposure. One study of German university students found people were especially likely to post online comments if they disagreed with an article or

if they wanted to persuade others in the comment sections. Moreover, the motivation to spread one's opinion more strongly predicts satisfaction with posting comments than other motivations, such as broadening one's knowledge or to simply wanting to discuss things with others (Springer, Engelmann & Pfaffinger, 2015). A separate field study of comment sections on a local news website predicted that incivility targeting political outgroups would increase as proportion of ingroup members increased, a prediction consistent with virtue signaling (Rains, Kenski, Coe, & Harwood, 2017). Contrary to their predictions, forum users became *less* likely to speak out against and derogate their ideological opponents as members of their ingroup grew in number. This pattern fits with coordinative goals, in which users are less motivated to undermine moral opponents if the moral consensus already favors the ingroup.

Other results interpreted as evidence of virtue signaling also have coordinative interpretations. For example, several independent studies have found that people are less likely to publicly condemn a transgressor or argue with an opponent once other people have already done so (Johnen et al., 2018; Sawaoka & Monin, 2018; Shugars & Beauchamp, 2019). Johnen and colleagues argue that people are not motivated to express outrage in these cases because it is more difficult to stand out from the crowd. Outrage becomes a less noteworthy signal of personal virtues when hundreds of others already expressed the same sentiment. While there is likely some truth to this explanation, condemning a transgressor also becomes less necessary once thousands have already punished it. From the coordinative perspective, seeing coalitions form behind an opposing rule or judgment is concerning. If consensus shifts to favor judgments opposite our own, the rules we prefer become ineffective guides for coordination. Consistent with this, people express more outrage towards transgressions committed by figures who have greater power to influence public opinion (Sawaoka & Monin, 2018). We use outrage and

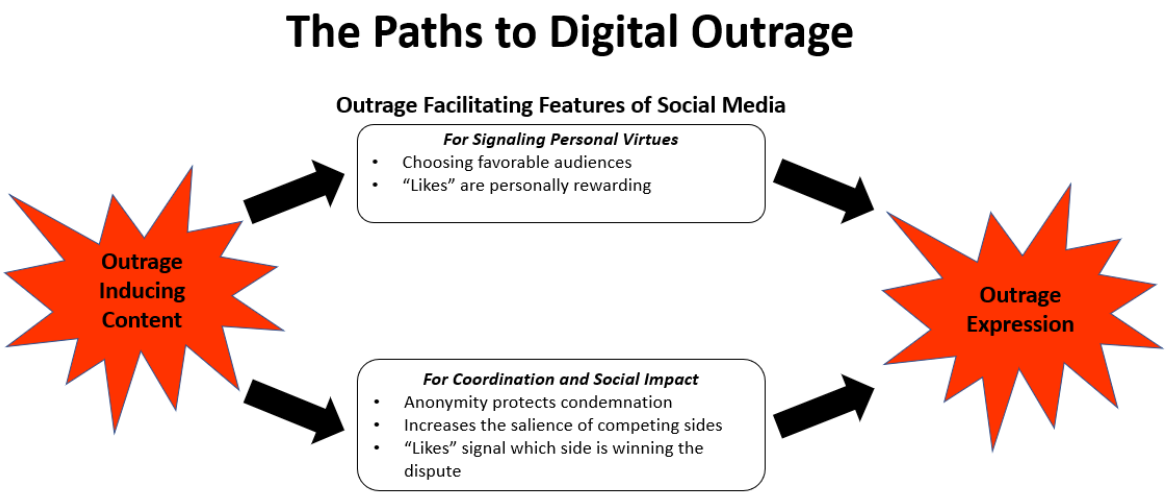


Figure 1. Social media facilitate outrage expression through multiple paths. First, social media increase exposure to potentially outrage inducing content. However, outrage is a tool with multiple functions. It can be used to improve our personal reputations, but it also coordinates people to our side in conflicts, helping our side win disputes and achieve its goals. Social media may facilitate both uses of outrage. They help use choose audiences that will reward our personal reputations most consistently. They also increase the salience of competing sides and broadcast a public "score" for who is winning the dispute—potentially transforming conversations into competitions. Thus, the present model proposes that digital outrage culture is a product of both people trying to signal personal virtues and have genuine impacts upon disputes.

shaming to combat opposing coalitions and to rally people to our side. But if the masses have already coordinated to condemn our opponents our job has already been done for us. When an opponent has already been publicly shamed on Twitter, we do not need to further punish them. At that point our side has already won.

Summary: Outrage to Cooperate, Outrage to Coordinate

Partner choice models and the side-taking perspective lay the theoretical groundwork for the virtue signaling and coordinative functions of digital outrage respectively. Both theories emphasize the importance of morality for garnering social support. Outrage in the service of virtue signaling conforms to prevailing norms to build personal reputation, making us more attractive cooperation partners. Outrage that aims to coordinate, however, drives onlookers

towards specific moral judgments, whether they be about moral rules, political issues, or specific people. The goal is to rally people behind a given issue or cause rather than attract cooperation partners. This suggest two different goals for outrage expression. *Cooperative outrag*e conforms to prevailing moral norms to build trust and rapport, leading to more successful cooperation. *Coordinative outrage* undermines opposing rules and people to coordinate others towards a cause that better serves our vested interests. Sometimes we condemn and shame others to show that we are trustworthy, sometimes we do so to ensure specific people or principles lose, gain, or maintain power.

OVERVIEW OF STUDIES

How did we get to a point where journalists and celebrities live in fear of an outrage mob descending upon them? Why do our Twitter and Facebook feeds have so many people yelling at each other? One explanation posits that people are trying to show off their moral character in front of their social networks. If they are the *most* outraged out of everyone then that will communicate just how much they believe in the cause. When everyone thinks like this, outrage mobs quickly spiral out of control. But outrage serves another purpose besides helping us fit in. Condemnation also coordinates people against our opponents in disputes. It robs celebrities and politicians we disagree with of influence and impels others to join our cause. Fighting over moral rules and issues is costly. It alienates friends and the benefits are not always clear. But digital media mitigates those risks. All the bickering and nastiness on social media may not just be people trying to amplify their personal reputation signal. Instead we may genuinely want to win disputes to enact social change, to coordinate behavior in ways relevant to our interests.

The present studies test two broad predictions that stem from the coordinative function of outrage. First, outrage should be felt and employed in response to seeing coalitions with opposing rules form. This motivates the use of outrage to protect the moral rules and people who help advance our vested interests. This leads to H1 through H3:

H1: Social media posts from an opposing political party with a high (vs low) number of likes/shares will elicit more outrage.

Furthermore, opposing posts that have more potential to spread should be especially likely to elicit outrage.

H2: Social media posts from an opposing political party with high (vs low) follower counts will elicit more outrage.

Lastly, witnessing others downvote an opposing coalition should reduce outrage.

H3: High (vs low) numbers of likes/shares for an opposing view will not increase outrage when they receive substantially more dislikes than likes.

Second, the motivation to undermine an opponent's reputation should produce greater use of condemnation and outrage than the motivation to receive personal reputation boosts. Moreover, people will feel especially free to invoke morality and outrage when the costs of expression are mitigated by anonymity.

H4: People instructed to write comments that will cause a moral transgressor to receive downvotes (vs. causing people to upvote your comment) will contain more anger and moral language.

H5: People's replies to a moral transgressor will contain more anger and moral language when anonymous than when identified.

I tested these hypotheses in three studies (see Figure 1 for an overview). Study 1 tested H1-H3 in a controlled, online survey using photoshopped webpages. Participants were shown 12 second animations of offensive tweets either accumulating a large amount or a very small number of retweets over their first 12 hours. Tweets came from accounts with either high or low numbers of followers, and some conditions suggested the tweets received substantial backlash. Subjective outrage and desire to respond were assessed following each animation. Study 2 tested H1 and H2 using real world interactions on Twitter in the wake of the Alabama abortion bill in May 2019. I

Overview of Studies

	Studies 1 and 2		Study 3
	The effects of virality upon outrage		The effects of explicit social coordinative vs. virtue signaling goals upon outrage
Study:	Study 1	Study 2	Study 3
Method:	Experimentally manipulate viral spread of ideologically opposing content	Measure viral spread of ideologically opposing and congruent content on Twitter	Manipulate: • Anonymity • Write comment to coordinate audience against target vs. to make audiences upvote self
Outcome:	Self-reported: • Outrage • Desire to Act	Behavioral Expressions: • Anger • Moral language	Self-reported: • Outrage Behavioral Expressions: • Outrage, moral language, mockery
Hypotheses:	H1, H2, & H3	H1 & H2	H4 & H5

Figure 2. Illustration of goals, methods, outcomes, and hypotheses tested in each study.

compared the anger and negative moral-emotional language of cross-ideological replies to Tweets with varying degrees of virality. Lastly, study 3 manipulated the hypothesized mechanism in Studies 1 and 2—that outrage is motivated by the goal of undermining opponents social support and influence. Study 3 tested H4 and H5 in an online survey of USF students that instructed them to either write comments that would make future USF participants upvote them personally vs. downvote a potential transgressor. Participants were led to believe their comments were either anonymous or identified and human raters scored each comment on outrage, moral language, and an exploratory variable, mockery.

STUDY 1

Method

Participants and Procedure

Procedures and analyses were pre-registered at https://osf.io/n2r7y/. A sample size of 240 participants was set to detect effect sizes of d = .4 between independent groups with power of .80. Participants were collected via Amazon's Mechanical Turk platform. To ensure familiarity with Twitter, people who did not have a Twitter account were excluded from participating using a brief screener survey. This resulted in an original sample of 245 participants. Five participants were excluded for failing a practice task requiring them to identify the number of likes and retweets in a screenshot of an example tweet (from the National Geographic Twitter account). This resulted in a final sample of 240 participants (M_{age} = 39.72, SD_{age} = 12.41, 59.2% women).

The main task of the survey showed participants animations of offensive tweets from their political outgroup gaining high or low amounts of likes/retweets over time (i.e., high versus low virality). At the beginning of the survey, participants indicated whether they leaned Republican or Democrat. Democrat participants were shown four profiles identifying as conservative Republicans and outrage inducing, right-wing tweets (e.g., "Murder and assault are spiraling out of control in cities all over the country thanks to 3rd world aliens who shouldn't even be here"). Republicans were shown profiles of liberal Democrats and outrage inducing, left-wing tweets (e.g., "Open borders, abolish ICE, citizenship for illegals, yes to all of it. Anything that stops this country from being run by a bunch of old white men").

Participants were also randomly assigned to see three types of pushback against each tweet. In the reply backlash condition, backlash was illustrated as it typically manifests on Twitter, through a high reply to like ratio. In other words, Tweets that are disapproved of by Twitter users tend to garner a high number of replies relative to the number of likes they receive (Roeder, Mehta, & Wezerek, 2017). In this condition, all Tweets accumulated roughly 5-10 times as many comments as likes. In the "downvote" condition participants were told to imagine that Twitter had a downvote button, and to respond as if the tweet received the number of downvotes displayed. Tweets accumulated as many downvotes in this condition as they did replies in the reply backlash condition. A third, control condition blurred out the portion of the tweet displaying the number of comments. See Appendices A-C for materials.

Participants saw screenshots of four, ostensibly real Twitter profiles and tweets. The number of followers each profile had and the number of favorites/retweets each tweet received was manipulated within subjects. Two accounts had relatively high follow counts (i.e., 50K – 80K) and two had relatively low follow counts (100 to 300). Participants were told we tracked several real tweets and screenshotted them every few hours. All demographic information was blurred out of the tweets and profiles, except for the bio that describes the user as liberal or conservative. Participants were then shown two offensive tweets that either 1) accumulated a high number of favorites/retweets (i.e., viral tweets with 7K-8K/1K – 2K) over 12 hours or a low number (i.e., 3-5/1-3) over 12 hours. Participants saw 5 screenshots of the page, labelled "hour 0" to "hour 12" at 3-hour intervals. At the "Hour 0" screen, participants were instructed to take a moment to read the tweet, then click the arrow button to see how many replies, retweets, and likes it accumulated over its first 12 hours. Each subsequent page (for the "3," "6", "9," and "12" hour marks) appeared on screen for 3 seconds. The manipulation of virality was independent of

the follower count manipulation. In other words, participants all saw tweets from two accounts with a large Twitter following (one going viral and one that did not go viral) and two tweets from accounts with small Twitter followings (one viral and one non-viral). Two items checked the success of the virality manipulation (i.e., "To what extent did the tweet gain support over time?" and "To what extent did this tweet influence people?"), four checked the perceived following size of each account ("How would you describe the size of this person's Twitter following?" 1 *Extremely small* – 5 *Extremely large),* and four checked the perceived ideology of each account (*Very liberal – Very Conservative*)

After each tweet participants indicated how outraged it made them feel using four items adapted from Tetlock et al., (2003) (i.e., angry, offended, outraged, upset; anchors = *Not at all - Very much*). As filler items, participants also completed two measures for how satisfied they felt (satisfied and pleased), two for fear (afraid and threatened), and two for surprise (surprised and caught off guard). Lastly, two items assessed subjective likelihood of commenting (i.e., "If you saw this on Twitter would you feel the need to speak up?" and "If you saw this on Twitter how likely would you be to write a reply?"). All items used 7 – point scales except where otherwise indicated.

Results

I first examined effects upon manipulation checks. A 2 (high vs low follower count) x 3 (reply backlash, vs downvote backlash vs control) mixed ANOVA (virality was not entered as a variable because it was manipulated after participants rated the following size of the accounts) indicated that the follow count manipulation was successful. Participants perceived the accounts with high followers as having a larger following (M = 3.73, SE = .05) than the perceived accounts with low followers (M = 2.03, SE = .04), $F(1, 237) = 898.30$, $p < .001$, $\eta_p^2 = .79$. The

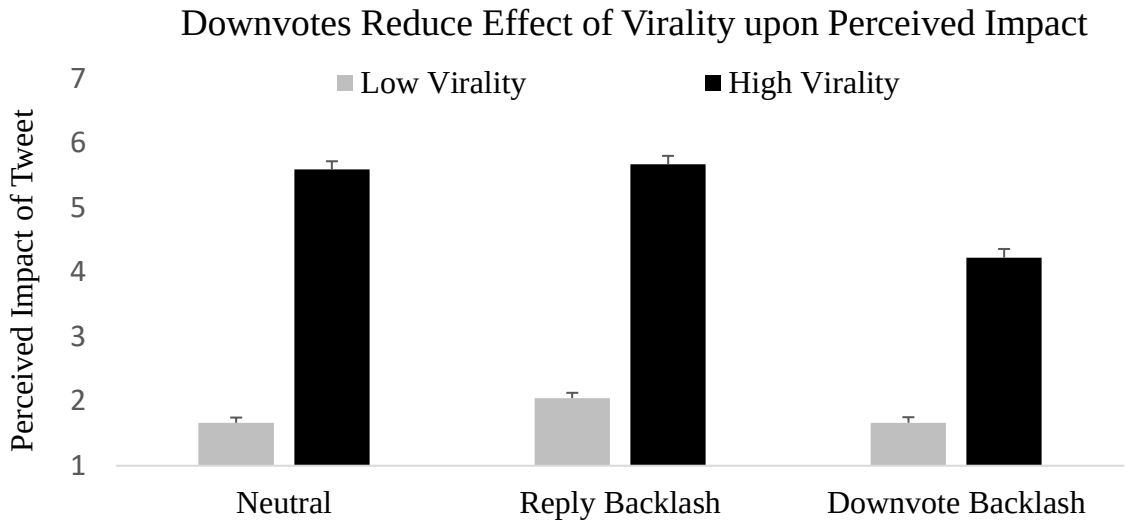


Figure 3. The positive effect of virality upon perceived impact is attenuated in the presence of hypothetical downvotes, but not by information that is present on Twitter (large numbers of replies).

follow count by backlash type interaction suggested that this difference was consistent across conditions, $F(1, 237) = 1.03$, $p = .36$, $\eta_p^2 = .01$. I next examined the perceived ideology of the Twitter accounts. Participants who saw offensive right-wing tweets (i.e., participants who indicated they leaned Democrat) perceived the offensive right-wing twitter accounts as conservative ($M = 6.27$, $SE = .07$) to an almost exactly equal degree that Republican leaning participants perceived the offensive left-wing tweeters as liberal ($M = 1.87$, $SE = .11$). A 2 (follower count) x 2 (lean Democrat vs lean Republican) x 3 (backlash type) mixed ANOVA suggested that this difference was consistent across accounts with high and low follower counts (follow count by participant ideology interaction: $F(1, 234) = .06$, $p = .80$, $\eta_p^2 < .001$). The three-way, follow count by participant ideology by backlash type interaction was also non-significant, $F(2, 234) = .86$, $p = .43$, $\eta_p^2 = .01$. Lastly, the virality manipulation was checked via a 2 (low vs high virality) x 2 (follow count) x 3 (backlash type) mixed ANOVA. As expected, high virality

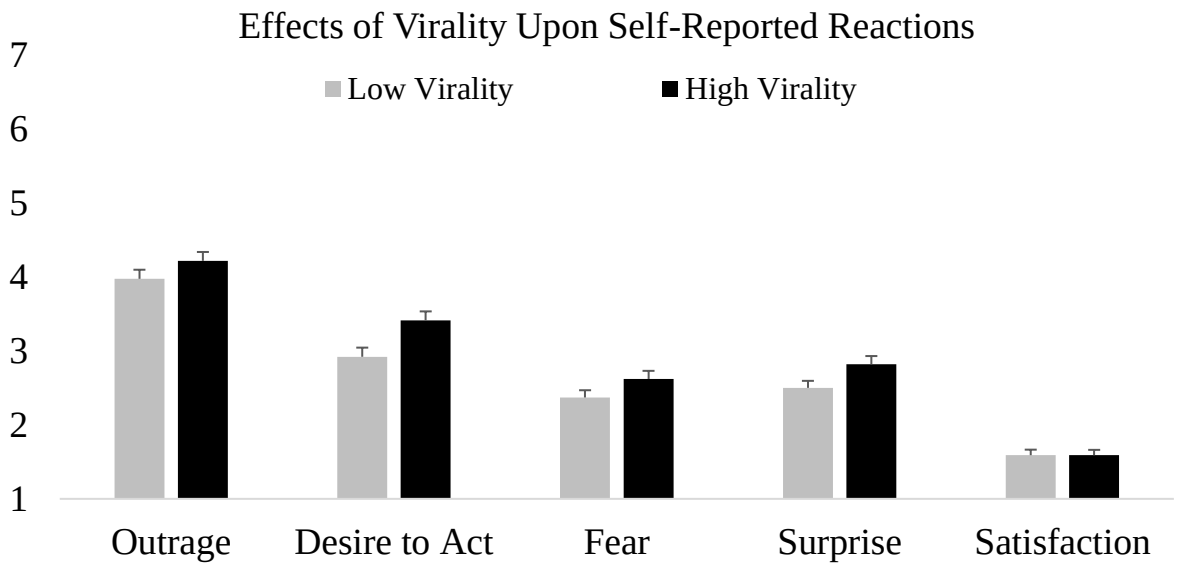


Figure 4. Effects of virality upon all self-reported reactions to politically opposing tweets.

tweets (M = 5.16, SE = .08) were perceived as substantially more supported and influential than low virality tweets (M = 1.79, SD = .05), $F(1, 237) = 1869.62$, $p < .001$, $\eta_p^2 = .89$. The virality by backlash type interaction was also significant $F(2, 237) = 27.56$, $p < .001$, $\eta_p^2 = .19$. The effect of virality upon perceived impact was substantially smaller in the downvote backlash conditions than the neutral or reply backlash conditions (see *Figure 2* for estimated marginal means).

To test Hypotheses 1-3, I conducted a 2 (follow count) x 2 (virality) x 2 (backlash type) mixed ANOVA with outrage as the dependent variable. Supporting H1, viral tweets evoked significantly more outrage than non-viral tweets (low virality: M = 3.97, SE = .12; high virality: M = 4.22, SE = .12, $F(1, 237) = 8.85$, $p = .003$, $\eta_p^2 = .04$. However, in contrast with H2, follow count did not impact outrage, $F(1, 237) = .009$, $p = .92$, $\eta_p^2 < .001$, nor did it moderate the effect of virality, $F(1, 237) = .02$, $p = .89$, $\eta_p^2 < .001$. The effect of virality also did not significantly differ across backlash type (virality by backlash type interaction: $F(2, 237) = .58$, $p = .56$, $\eta_p^2 =$

.01); thus, H3 was also not supported. However, examining simple effects revealed suggestive evidence of virality affecting outrage differently in the control vs downvote backlash conditions. When no information of backlash was provided, the effect of virality trended in the low follower (ΔM = .32, SE = .19, p = .08), not the high follower conditions (ΔM = .03, SE = .20, p = .89). But when tweets were accompanied by downvotes, the effect of virality trended in the high follower conditions (ΔM = .36, SE = .21, p = .08), not the low follower conditions (ΔM = .01, SE = .19, p = .95). Furthermore, an exploratory mixed ANOVA, dropping the reply backlash condition, produced a marginal follow count by virality by backlash type interaction, $F(1, 237) = 3.09$, p = .08, η_p^2 = .02. Thus the effect of virality may differ slightly across conditions, but the present evidence of this is merely suggestive.

Effects upon an alternative outcome, desire to act, followed a similar pattern to results for outrage but had slightly stronger effects. Viral tweets produced a significantly stronger desire to act (M = 3.41, SE = .12) than non-viral tweets (M = 2.92, SE = .12), $F(1, 237) = 28.41$, $p < .001$, η_p^2 = .11. However, as with outrage, follower count of the offensive tweeter did not impact the desire to act $F(1, 237) = .01$, p = .91, $\eta_p^2 < .001$, nor did it moderate the effects of virality $F(1, 237) = 2.82$, p = .16, η_p^2 = .01. Examining effects upon the remaining filler, self-reported feelings revealed similar effects of virality upon fear (ΔM = .25, SE = .07, p = .001) and surprise (ΔM = .31, SE = .08, $p < .001$), but not satisfaction (ΔM = .001, SE = .07, p = .99)

Lastly, an exploratory analysis examined differences across participants who leaned Republican vs those who leaned Democrat. Results from the 2 (follow count) x 2 (virality) x 3 (backlash type) x 2 (Democrat vs Republican) mixed ANOVA revealed a significant three-way interaction between follow count, virality, and participant ideology, $F(1, 234) = 4.32$, p = .04, η_p^2 = .02. For Democrats, the viral tweets evoked more outrage (relative to non-viral tweets) when

the tweet author had a large Twitter following (ΔM = .44, SE = .14, p = .001) and a small Twitter following (ΔM = .24, SE = .13, p = .07). For Republicans, the effect of virality trended in a positive direction for accounts with small followings (ΔM = .23, SE = .20, p = .20) and in a negative direction for accounts with large followings (ΔM = -.28, SE = .21, p = .21). In sum, the effect of virality was more consistent in participants who leaned Democrat. Dropping people who leaned Republican increased the effect size of virality upon outrage from η_p^2 = .04 to η_p^2 = 07. Notably, the sample contained fewer Republicans (n = 71) than Democrats (n = 169).

Discussion

In Study 1, tweets from political opponents that accumulated high numbers of likes and retweets (i.e., high virality) generated significantly more outrage and the desire to reply than tweets that accumulated low number of likes and retweets (supporting H1). This effect persisted even when the offensive tweets received large numbers of downvotes. Apparent backlash, in the form of downvotes, did make the viral tweets seem less impactful. However, this did not appear to reduce participants' outrage or desire to respond (in contrast with H3). Furthermore, the number of followers each account had did not impact outrage or desire to reply, nor did it moderate the effects of virality upon outrage (in contrast with H3). Effects were also more consistent among participants who leaned Democrat than those who leaned Republican (in fact dropping the latter from analyzes nearly doubled the effect size of virality upon outrage). In sum, Study 1 provided evidence of a small effect of virality upon both felt outrage and the desire to reply, but little evidence of an effect of following size or apparent backlash.

STUDY 2

Study 2 aimed to replicate the effect of virality upon outrage in real world conversations on Twitter. Rather than measuring self-reported outrage, Study 2 employs a dictionary-based method to detect the presence of anger and moral emotional language in cross ideological and ideological homogenous interactions. Thus Study 2 builds upon Study 1 by testing the effects of virality (H1) and actual following size (H2) upon moral emotional language in real world behavior and by comparing interactions between ideological similar (e.g., conservatives with conservatives) and ideologically dissimilar (i.e., conservatives with liberals) pairs of users.

Method

Data Collection

Tweets containing the word "abortion" were collected using the Twitter streaming API in the wake of the Alabama abortion bill, passed on May 14th of 2019. Tweets were collected on May 16th. First, only top-level tweets (i.e., tweets that are not replies to other tweets) were retained, resulting in an original corpus of 153,412 tweets. To narrow the sample down to comment threads more likely to contain at least one cross-ideological interaction, tweets with fewer than 3 replies were excluded. The overwhelming majority of tweets had no replies (n = 109,160) or only 1 reply (n = 31,824; many of which were users replying to themselves). Thus this step alone substantially reduced the size of the corpus of top-level tweets (5,676 top-level tweets in total). Over the following week I used the Twitter search API to collect replies to users in the data set from May 16th to the 18th. I then matched replies to their corresponding top-level

tweets, excluding tweets that were replies to other replies in each tweet thread (a thread refers to a top-level tweet, all of its direct replies, and replies to other replies within the thread). If a user replied more than once to the same tweet, I only kept their first reply. This resulted in an original sample of 84,190 direct replies to top-level tweets. Lastly, because I was most interested in how people responded to top-level tweets, I only retained the first 30 replies to minimize how other replies may have impacted the conversation over time. The final corpus consisted of 5,676 tweets and 44,215 replies.

Estimating Political Ideology of Twitter Accounts

The ideology of user accounts was estimated using a previously validated computational model (Barbera, 2015; Barbera, Jost, Nagler, Tucker, & Bonneau, 2015), implemented in the "tweetscores" package in R. For example, previous validation studies have shown ideology estimates derived from this model predict real world political party registration with over 90% accuracy (Barbera, 2015). The model infers users' political ideology based on the assumption that people prefer to follow politicians with similar ideologies of their own. In other words, the accounts someone choose to follow on Twitter contain information about their personal ideology. Barbera's (2015) original paper contained two stages for estimating ideology, one estimating the political ideology of a set of political elites and a second stage estimating the ideology of 32 million Twitter users across six countries based on the political elites they follow. One limitation of this original method was that users had to follow at least one political elite to estimate their ideology, leading to substantial loss of data. A later paper (Barbera et al., 2015) expanded the list of followed accounts used to estimate ideology from political elites (e.g., congress people and presidents) to accounts commonly followed by liberals and conservatives (e.g., Stephen Colbert and Rush Limbaugh). Ideology scores for this expanded list of political

accounts are normalized to follow a normal distribution with a mean of zero and standard deviation of one. The pre-estimated ideology of accounts in the tweetscores package, used to estimate the ideology of Twitter users in the present data, were last updated in October of 2018. Lists of followed accounts for all Twitter accounts in the present data were collected through the Twitter API from December 2019 to January 2020. Ideology estimates were obtained using the "estimateIdeology2()" function in the tweetscores R package (which implements the ideology estimation procedure using the expanded list of political accounts described in Barbera et al., 2015).

After obtaining ideology scores for tweets and replies, replies were classified as ideologically cross-cutting if the ideology estimate of the replying user had the opposite sign as the ideology of the user who posted the original tweet (e.g., a user with an ideology of -.5 replying to a user with an ideology of .5 represented a liberal replying to a conservative). Replies were classified as ideologically homogeneous if the replying user and original tweeter had ideology estimates of the same sign. Ideology estimates were successfully obtained for 7,363 cross-cutting replies and 14,729 homogeneous replies.

Language Analysis

All tweets were pre-processed by removing URLs, emoticons, hashtags, punctuation, numbers, and extra white space. Only Tweets with 10 or more words were included in analyses. The text of each tweet was analyzed using a dictionary-based approach following Brady and colleagues' (2017) similar analysis of moral-emotional language on Twitter. All tweets and replies were scored on anger and moral language by calculating the proportion of each tweet composed of words from validated dictionaries (anger dictionary: n = 329 words; negative moral-emotional dictionary: n = 38; positive moral-emotional: n = 20; Brady et al., 2017;

Graham, Haidt, & Nosek, 2009; Pennebaker, Boyd, Jordan, & Blackburn, 2015). For example, a reply reading "I hate Trump!!! #impeach" would be split into four words "I," "hate," "Trump," and "impeach" and assigned an anger score of .25 because one word (hate) out of four is in the anger dictionary. Primary analyses focus on results for anger and negative moral-emotional language because they map most directly onto moral outrage. Dictionaries for negative and positive moral emotional language were taken from Brady et al., (2017), who constructed them from overlapping words in validated dictionaries of negative affect, positive affect, and moral words. The moral language dictionary is located in Appendix E (note that the full anger dictionary from LIWC is not included in appendices because it is proprietary).

Predictors

Predictor variables included the number of followers of authors of top-level tweets, the number of likes and retweets received by each tweet top-level tweet, and the number of replies to each tweet top-level tweet. A virality score was calculated by standardizing and averaging the number of retweets and likes.

Results

To mitigate the influence of a small number of highly viral tweets in the data set, top-level tweets with virality scores three standard deviations above the mean were removed from the data set. Follower counts were rescaled by taking the natural logarithm to adjust for the extremely wide range of followers (0 to 45,196,481).

While some replies in the data set were not clustered with other tweets (for example, 24% of top-level tweets only had a single reply and 44% had two), many replies were nested within the same top-level tweet. To check whether multi-level modeling was necessary to account for potential non-independence, random effects only models were estimated for anger, negative

moral-emotional, and positive moral-emotional language scores, entering only a tweet identification number as a random effect. Intra-class correlations for models predicting anger (ICC = .05), negative moral emotional (ICC = .05), and positive moral emotional words (ICC = .02) were all low, suggesting that multi-level models to account for non-independence were not necessary (Dyer, Hanges, & Hall, 2005; Hox, 2002). Thus for the primary analyses I report the results of linear regressions.[1]

Regressions predicting anger, negative moral-emotional language, and positive moral-emotional language within replies were run in three steps. The first step entered virality score of the top-level tweet, the log transformed number of followers of the account belonging to the top-level tweet, and whether the reply was cross ideological or ideologically homogeneous. The second step entered all three two-way interactions, and the third step entered the three-way interaction. Detailed model results are provided in Table 2.

In support of Hypothesis 1, the effect of virality upon anger words and negative moral emotional language was moderated by whether the reply was cross-cutting or homogeneous (interaction effects for anger and negative moral emotional language respectively: b =-.42, p = .003 and b =-.30, p = .004). In cross cutting interactions, more viral tweets were targeted with angrier (b = .44, p < .001) and negatively moral (b = .28, p < .001) replies. No such relationship existed in homogeneous interactions (b = .02, p = .84 and b = -.02, p = .77 for anger and negative moral words respectively). The interaction between log transformed follower count and cross-cutting (vs homogeneous) was also significant for anger (b = .03, p = .03) and negative moral language (b = .04, p < .001). However, the simple effects did not follow the pattern predicted by

[1] As a robustness check, I estimated multi-level models for all key results, entering tweet ID as a random effect and predictors (virality score, follower count, and cross-cutting vs homogeneous) as level 2 fixed effects. No results differed substantively across this analytic approach and the linear regressions reported in the primary analyses.

Table 1. Relationships between tweet metadata and anger/moral-emotional language

Outcome	Predictor	*b* (SE)	*p*	Model statistics
Anger	Virality	0.17 (0.7)	.011*	R^2. = 0004, $F(3,21875) = 2.68$, $p = .05$
	Follower count	-.003 (0.01)	.652	
	Reply type	-0.02 (0.01)	.232	
	Virality x Followers	-0.02 (.02)	.465	R^2. = 0010, $F(6,21873) = 3.68$, $p = .001$
	Virality x Reply type	-0.42 (0.14)	.003*	
	Followers x Reply type	.03 (0.02)	.027*	
	Virality x Followers x Reply type	-0.07 (0.04)	.077	$R^2 = .0012$, $F(7,21871) = 3.60$, $p < .001$
Negative Moral	Virality	0.10 (0.05)	.052	R^2. = 0006, $F(3,21875) = 4.02$, $p = .007$
	Follower count	.01 (0.01)	.065	
	Reply type	-0.06 (0.03)	.033	
	Virality x Followers	-0.02 (0.02)	.244	R^2. = 0017, $F(6,21873) = 6.17$, $p < .001$
	Virality x Reply type	-0.30 (0.10)	.004*	
	Followers x Reply type	.04 (0.01)	< .001*	
	Virality x Followers x Reply type	-0.06 (0.03)	.063	$R^2 = .0018$, $F(7,21871) = 5.78$, $p < .001$
Positive Moral	Virality	0.01 (0.03)	.723	R^2. < 0001, $F(3,21875) = .50$, $p = .70$
	Follower count	-0.003 (0.004)	.423	
	Reply type	.02 (0.03)	.389	
	Virality x Followers	-0.02 (0.01)	.138	R^2. = .0002, $F(6,21872) = .85$, $p = .53$
	Virality x Reply type	.09 (0.07)	.206	
	Followers x Reply type	<.001 (0.1)	.994	
	Virality x Followers x Reply type	-0.01 (0.22)	.654	$R^2 = .0002$, $F(7,21871) = .76$, $p = .62$

Note: Model statistics are provided at three steps: 1) predictors entered with no interactions 2) predictors with all 2-way interactions 3) predictors, two-way interactions, and the three-way interaction.

Hypothesis 2. In cross-cutting interactions, log follower count was associated with marginally less anger ($b = -.02$, $p = .05$) and negative moral language ($b = -.01$, $p = .11$). But follower count was positively associated with negative moral language ($b = .03$, $p < .001$) in homogenous interactions (and not associated with anger, $b = .01$, $p = .27$). No effects were observed upon positive moral language (see Table 2 for detailed results).

Three-way interactions between virality score, log transformed follower count, and cross-cutting (vs homogeneous) were marginal for anger (*b* -.07, *p* = .08) and negative moral language (*b* -.06, *p* = .06). Interestingly, the virality by reply type (cross-cutting vs homogeneous) were slightly exaggerated at high follower counts (effect of virality among cross-cutting replies: *b* .55, *p* < .001; among homogeneous: *b* -.12, *p* = .31) compared to low follower counts (effect of virality among cross-cutting replies: *b* = .38, *p* = .004; among homogeneous: *b* = .12, *p* = .21) for anger. Negative moral language exhibited this same pattern. The interaction between virality and reply type was slightly more pronounced at high follower counts (effect of virality among cross-cutting replies: *b* .35, *p* = .003; among homogeneous: *b* -.14, *p* = .11) than low follower counts (effect of virality among cross-cutting replies: *b* = .31, *p* < .001; among homogeneous: *b* .09, *p* = .66).

Secondary Analyses

Outrage Amplifying Effects of Virality. Despite excluding tweets with fewer than 10 words, a substantial number of replies contained no words in the anger or negative moral emotional word dictionaries. This produced substantial clustering at zero in both primary outcome variables. As a secondary analysis, I re-ran the same regressions excluding tweets with no anger and no negative moral emotional words, reducing the sample of replies to 6,595 and 3,888 in the regressions predicting anger and negative moral words respectively. In other words, these analyses only include tweets exhibiting some degree of anger or moral language. Thus these tests may represent effects of virality upon outrage *amplification* rather than *origination*.

Using this smaller sample, the two-way interactions between virality and reply type no longer had significant effects upon anger (*b* = -.37, *p* = .19) or negative moral emotional language (*b* = -.51, *p* = .13). Likewise, the interactions between follower count and reply type no

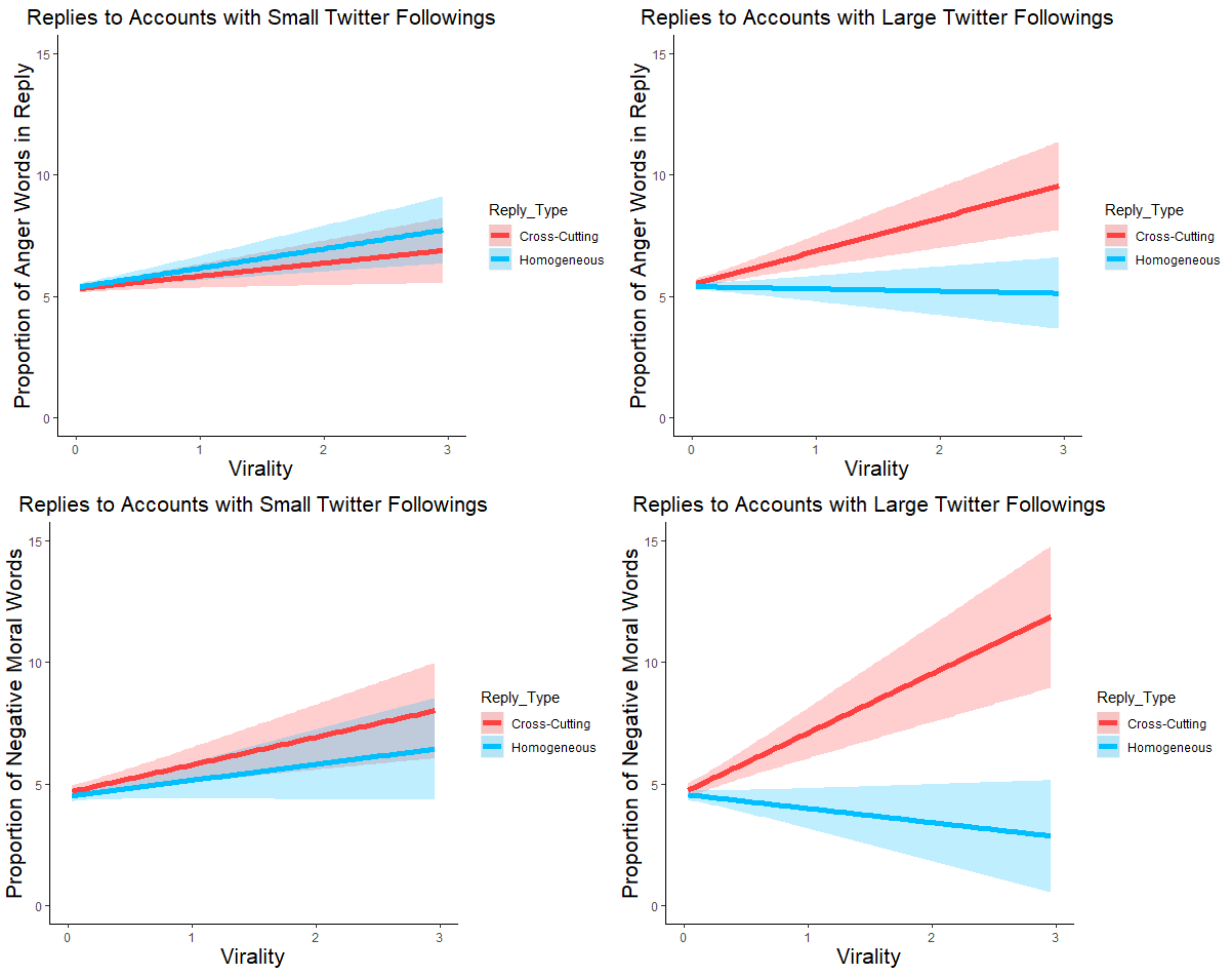


Figure 5. Relationship between virality and moral-emotional language in ideologically cross-cutting and homogeneous replies. Small and large Twitter followings represent effects estimated at one standard deviation below and above the mean log transformed follower count. Scores for anger and moral language are proportion of words in tweets from validated dictionaries for moral and negative moral-emotional language. Data excludes replies with zero anger and negative moral-emotional words.

longer affected anger (b = 0.01, p = .64) or negative moral words (b = .03, p = .39). However, the effects of the three-way interactions were substantially larger in this sample. The final model predicting anger (i.e., containing all three predictors, two-way interactions, and the three-way interaction) was significant, $F(7, 6563) = 5.16$, $p < .001$, $R^2 = .01$, as was the three-way interaction, b = -.32, SE = .08, p = < .001. Probing the interaction revealed that the hypothesized interaction between virality and reply type only emerged among replies to accounts with large

twitter followings. For accounts with relatively low numbers of followers, virality predicted increases in anger among both cross ideological replies (b = .52, p = .02) and homogenous replies (b = .81, p < .001). However, at high follower counts, not only was the effect of virality among cross-cutting replies substantially larger (b = 1.37, p < .001), virality no longer predicted anger among homogeneous replies (b = -.09, p = .70). Results for negative moral emotional language followed a similar pattern. Again, the final step of the model was significant, F(7, 3,865) = 4.508, p < .001, R^2 = .01, as was the three-way interaction, b = -.35, SE = .10, p = < .001. At low follower-counts, virality did not predict increased use of negative moral language among either cross-cutting (b = .33, p = .21) or homogeneous replies (b = .49, p = .08). However, at high follower counts, the predicted interaction emerged. Among cross-cutting replies, virality predicted greater user of negative moral emotional language (b = 1.55, p < .001), while there was no effect of virality among homogenous replies (b = -.32, p = .31) (see Figure 3 for visualizations of interactions).

Negative Binomial Models of Word Counts. The previous analyses calculated the proportion of anger and negative moral emotional within in each tweet (i.e., by dividing the number of words from each dictionary by the total number of words in the corresponding tweet). While this approach is common (for example, it is the default output of the LIWC software) and has been previously applied to tweets (Eichstaedt et al., 2015; Jordan, Pennebaker, & Ehrig, 2018), other approaches analyze word counts without calculating proportions (e.g., Brady, et al. 2017). Both methods have their own limitations. Proportions control for the fact that longer tweets are more likely to contain words from the given dictionary by virtue of their length. However, they may also underestimate a sentiment, such as anger, among twitter users who tend to write longer sentences. As a robustness check, I also present analyses of word counts.

Furthermore, analyzing count data allows for the use of discrete probability distributions, such as the negative binomial, that are effective at modeling the skewed, zero-inflated characteristics of the present outcome variables.

I first estimated Poisson models to test for the presence of overdispersion (i.e., whether the standard deviation of the outcome variable substantially exceeded its mean). The Poisson distribution assumes that the mean and variance are equal. If this assumption is violated, then negative binomial regression is appropriate. Using the AER package in R to implement the test of overdispersion describe by Cameron and Trivedi (1990), both anger (α = .32, z = 14.86, p < .001) and negative moral-emotional (α = .37, z = 12.48, p < .001) word counts exhibited significant overdispersion, thus negative binomial regressions were used to model word counts.

Results were similar to those analyzing proportions. Type of reply moderated the effects of virality upon anger words, b = -.24, SE = .08, z = 2.61, p = .002. In cross-ideological replies, virality predicted significantly more anger words, b = .21, SE = .06, z = 3.63, p < .001. On average, a tweet with a virality score of three was predicted to contain twice as many anger words than tweets with a virality score of zero (see *Figure 4*). No such effect was observed in homogenous replies, b = -.02, SE = .05, z = -.27, p =.64. Reply type also moderated the relationship between virality and number of negative moral-emotional words, b = -.30, SE = .10, z = -2.86, p = .004. Virality predicted negative moral words in cross-cutting replies, b = .26, SE = .08, z = 3.27, p < .001, but not homogeneous replies, b = -.04, SE = .07, z = -.61, p = .54. Similar to the results for anger, the predicted number of negative moral-emotional words in replies to tweets with a virality score of three was twice that of tweets with virality scores of zero. Neither the three-way interaction predicting anger, b = -.02, SE = .02, z = -.85, p = .39, or negative moral words was significant, b = -.05, SE = .03, z = -1.56, p = .12.

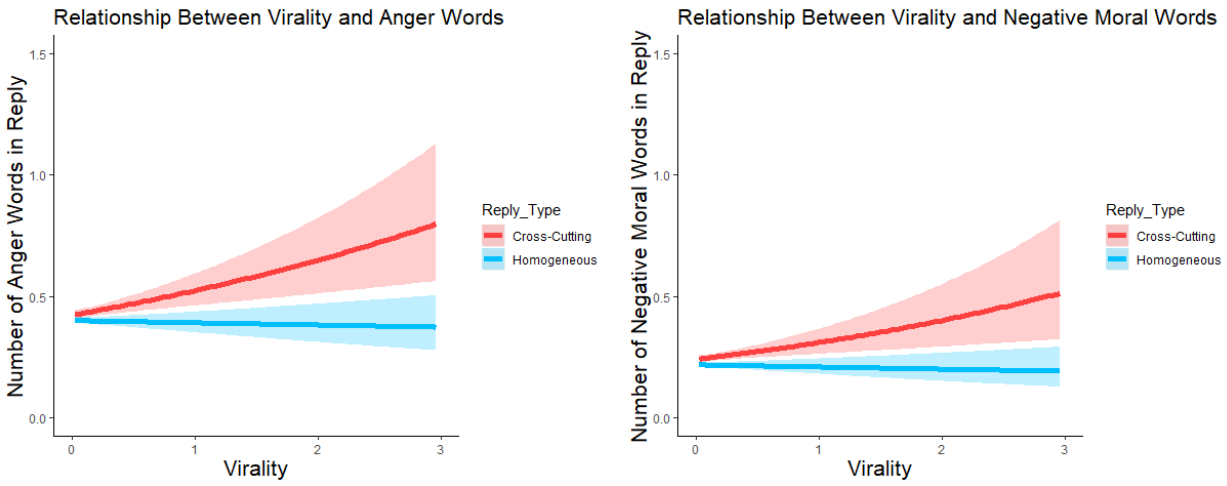


Figure 6. Relationship between virality and moral emotional language word counts modeled via negative binomial regression.

Discussion

Study 2 successfully replicated the effect of virality upon outrage in real world behavior on Twitter. Replies to political outgroups on Twitter were significantly angrier and used more uniquely moral, negative emotion words for viral than non-viral tweets (supporting H1). This effect was robust to multiple statistical approaches. Moreover, a secondary analysis focusing on moral outrage amplification rather than origination, revealed not only a substantially larger effect of virality upon angry and moral emotional language, but a moderating role of following size. Among accounts with few followers, the effect of virality upon anger and negative moral language was relatively small and did not differ across homogeneous and cross-cutting replies. However, viral posts from accounts with large followings evoked significantly more outrage and this effect only emerged in ideologically cross-cutting replies. While this may not provide direct support for the hypothesized positive relationship between following size and outrage (H2), following size may play an important role in determining when viral tweets evoke outrage. When users with small followings happen to go viral, they may attract relatively little negative, moral-emotional language. But when hugely influential political opponents post on moral topics, the

threat of virality may loom especially large, prompting the use of specifically moral, negative emotional language to combat the spread of opposing ideas.

STUDY 3

Studies 1 and 2 found that people feel and express outrage in response to viral content from political outgroups. I theorize that outrage is both felt in response to seeing opposing ideas gain favor and expressed to combat the spread of those ideas. One limitation of Studies 1 and 2 is that they do not distinguish between subjective feelings of outrage and behavioral expressions of outrage. It is possible, for example, that people use outrage strategically as a tool to suppress opposing ideas or to signal their personal virtues without feeling outraged subjectively. Study 3 has three goals. First, it tests whether outrage expressions increase when people have an explicit goal of coordinating against an opponent, even in the absence of shifts subjectively felt outrage. Second, it tests whether the explicit motivation to coordinate people against an opponent entails greater uses of outrage than the explicit motivation to personally gain social rewards on social media (e.g., likes or upvotes) (a test of Hypothesis 4). Third, I test whether anonymous (vs identified) digital contexts make people feel even more free to express outrage when trying to prevent the spread of opposing ideas.

Method

Participants and Procedure

Again, a sample size of 320 was set to detect effect sizes of d = .4 between independent groups in the design with power of .80. Data collection fell short of this goal, with an initial total of 217 participants (M_{age} = 20.43, 180 Women, 36 Men, and 1 who selected other but did not identify their gender) collected from the USF psychology participant pool. Participants were told

they were taking part in a study of online interactions in which they would read comments from other USF students, write replies, and upvote/downvote other responses. Half of participants were instructed to write comments they thought would increase the number of downvotes another person's comment to received downvotes; half wrote comments they though would cause them to receive upvotes themselves. See instructions below (text altered between conditions is underlined with differences in brackets):

> "We are studying how effective people are at manipulating the number of likes or upvotes they receive on social media. We've collected some comments from other USF students about campus life. We want you to write some replies for us to show participants in a future study, and we will test whether people upvote your comments [your comments make the person you are replying to receive more downvotes]. In other words, on the following pages we want you to write replies that you think will make people upvote your comments [cause people to downvote the person you are replying to]."

Participants were then randomly assigned to read that their replies will be anonymous vs. identified:

> "Below is an image of how your comment will appear to future participants. USF students participating in our future studies will read the replies you write as depicted below. It is important for you to know that your real name will appear next to the comments that you write, making your identity known to future participants who read whatever replies you write [It is important for you to know that we will collect no data about your identity. Whatever replies you write will be completely anonymous and your identity will be unknown to future participants]."

Below these instructions was a screenshot of a photoshopped comment section, illustrating how their comment would appear. Each prompt was following an attention check. To ensure participants understood the task, participants first had to summarize, in their own words, the goal of the comment they were supposed to write. Response were coded as "pass" if they correctly summarized the goal of their comment (to make others upvote you vs. make others downvote someone else) and coded as "fail" otherwise. The second attention check (for anonymity) asked participants to indicate what would appear next to their comment. Choices included, "Your real name," "A profile picture," "Your year in school," and "There will be no identifying information whatsoever." All participants who failed at least one attention check were excluded from analysis (see Appendix F for full manipulation materials).

Participants read two comments about "USF life," ostensibly written by other students. To help participants get acquainted with the task and to mask the purpose of the study, participants first read a neutral comment designed to elicit minimal outrage reading, "I like the set up for the gym a lot but oh my god does it get crowded in there. Have to stand around awkwardly for 10 mins just to get a bench some days." The second, offensive comment, gave students the opportunity to express moral outrage, "Wow was not expecting for there to be soooooo many black people here. Nothing wrong with that just not used to it. Weird." See Appendix G for comment materials.

Measures

After reading each comment, participants were asked if they wanted to 1 – upvote, 2 – downvote, or 3 – do nothing. All participants were then asked to write a reply to each comment. The primary dependent variables were outrage and moral conviction expressed in participants' replies to the offensive comment. Two independent raters (both of whom identified as White)

each coded comments for outrage, moral conviction, and a third, exploratory variable, mockery. Coding instructions contained the original comments to which participants responded. Outrage was described to raters as a mix of condemnation and emotions such as anger and disgust (Skitka, 2010; Tetlock et al., 2003). Coders rated outrage on a four-point scale (0 – Not at all outraged, to 3 – a strong degree of outrage). Perceived moral language was coded on an item adapted from Skitka's (2010) measure of moral conviction (i.e., The commenter expresses their moral convictions, 0 – No at all to 3 – The comment seemed strongly tied to their moral beliefs and convictions). Lastly, raters coded for mockery (i.e., The commenter mocks the original comment 0 – Not at all to 3 – The entire comment is blatantly parodying or making fun of the other user). Full coding instructions are provided in Appendix H.

Following guidelines from Hallgren (2012), inter-rater reliability was assessed via intra-class correlations in the *irr* package in R. I specified a two-way model (since both raters rated all comments, i.e., the design was fully crossed) focusing on consistency (rather than absolute agreement), and indicating that the final measures used for outrage, moral language, and mockery were the means of both coders rating. ICCs for outrage (.88) and mockery (.90) indicated strong agreement, while the ICC for moral conviction showed only moderate reliability (.68).

Lastly, participants completed the same self-report measure of subjective outrage as in Study 1.

Results

All participants who failed either manipulation check or upvoted the offensive comment were excluded from analyses. This reduced the sample to 150 participants (M_{age} = 20.51, 129 Women, 20 Men, and 1 who selected other but did not identify their gender). Exclusions and

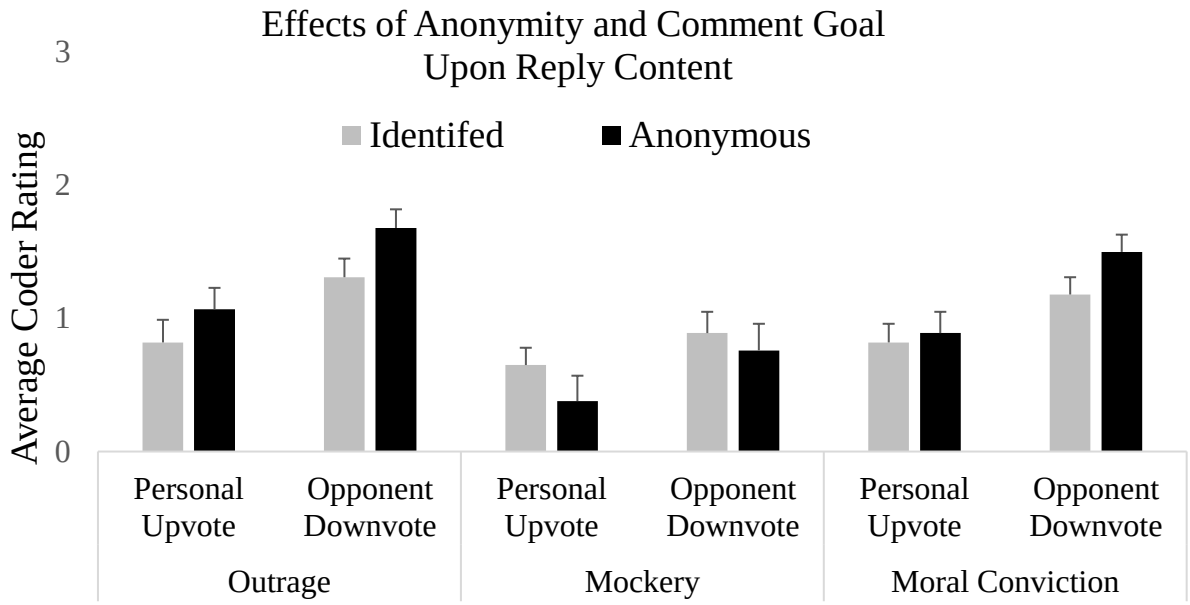


Figure 7: Effects of anonymity and comment goal upon each type of coder rated content.

analyses were pre-registered at https://osf.io/n2r7y/. The final sample had .80 power to detect η_p^2 = .05. A 2 (goal of comment: opponent downvote vs personal upvote goal) by 2 (anonymous vs identified) ANOVA, entering outrage as the dependent variable, found a significant main effect of comment goal, $F(1,146) = 13.12$, $p < .001$, $\eta_p^2 = .08$ (supporting H4), and a marginal effect of anonymity, $F(1,146) = 2.82$, $p = .07$, $\eta_p^2 = .02$ (partial support for H5). Participants with the explicit goal of making someone else receive downvotes expressed significantly greater outrage ($M = 1.49$, $SD = .88$) than participants who tried to receive upvotes ($M = .96$, $SD = .99$). Anonymous participants ($M = 1.37$, $SD = 1.02$) expressed marginally more outrage than identified participants ($M = 1.10$, $SD = .90$) The effect of comment goal was not moderated by anonymity $F(1,146) = .014$, $p = .71$, $\eta_p^2 < .001$. Examining expression of moral conviction, participants who tried to inspire downvotes expressed more moral conviction ($M = 1.34$, $SD = .83$) than participants who tried to receive upvotes ($M = .86$, $SD = .88$), $F(1,146) = 12.29$, $p < .001$, $\eta_p^2 = .08$. Neither anonymity, $F(1,146) = 1.46$, $p = .23$, $\eta_p^2 = .01$, nor the anonymity by

comment goal interaction affected expressed conviction, $F(1,146) = .86$, $p = .35$, $\eta_p^2 = .01$. Lastly, participants in the downvote condition ($M = .82$, $SD = 1.15$) used mockery ($M = .49$, $SD = .91$) marginally more often than participants in the upvote condition, $F(1,146) = 2.82$, $p = .07$, $\eta_p^2 = .02$. Neither anonymity, $F(1,145) = 1.68$, $p = .20$, $\eta_p^2 = .01$, nor the anonymity by comment goal interaction affected mockery, $F(1,145) = .17$, $p = .68$, $\eta_p^2 = < .001$.

Comparing expressed outrage with self-reported outrage found no relationship, $r(148) = .11$, $p = .17$. Comment goal, $F(1,144) = .03$, $p = .86$, $\eta_p^2 < .001$, anonymity $F(1,144) = .16$, $p = .69$, $\eta_p^2 < .001$, and the interaction, $F(1,144) = .04$, $p = .85$, $\eta_p^2 < .001$, all had no effect on self-reported outrage. While participants expressed substantially greater outrage and moral conviction in the downvote conditions, they did not feel any more outrage subjectively nor did outrage expression correlate with feeling outrage overall.

Discussion

When people explicitly try to coordinate others against someone they disagree with, they express substantially greater outrage and moral conviction than when they try to secure upvotes for themselves. Anonymity may further increase outrage expression, but the effect of anonymity was small and marginally significant. Expressions of outrage shifted substantially despite no change in subjective outrage. When people have explicit goals to manipulate public opinion, they are more likely to express outrage even if in the absence of increased outrage. I also found suggestive evidence that mockery provides an alternative tool for coordinating audiences against targets.

These results do not preclude outrage motivated by implicit goals to improve personal reputation (Jordan & Rand, 2019), but they do suggest when people explicitly try to receive social rewards online, outrage is not their primary strategy. This is consistent with Jordan and

Rand's (2019) finding that people only use outrage to signal personal reputation when pro-social behavior is not an option. However, the present results suggest that improving personal reputation (at least in an absolute sense) may not be the primary drive behind digital outrage, as efforts to coordinate onlookers against an opponent evokes relatively more outrage expression.

Notably, ratings of moral conviction suffered from relatively low reliability. This is perhaps unsurprising given that moral conviction researchers argue self-report is the most appropriate method for assessing moral convictions (Skitka, 2010). In other words, when someone is criticizing and distancing themselves from another, it can be difficult to tell whether they are driven by moral concern versus dislike without asking them. Thus results for moral conviction should be interpreted with caution. Lastly, while subjective and expressed outrage did not correlate overall, among people who expressed any degree of outrage (i.e., people who either rater score as a 1 or higher on outrage), both outrage indicators correlated at $r(110) = .26$, $p = .006$. Thus, among people who expressed outrage, feeling outrage more intensely predicted more extreme expressions of outrage.

GENERAL DISCUSSION

Common sense suggests that offensive ideas trigger outrage because of their content. However, the present studies provide evidence that digital outrage is not only about the offensive content we see; it is about the potential for that content to spread and gain influence. In Study 1, animations of offensive tweets going viral (compared to similarly offensive, non-viral tweets) triggered subjective outrage and an increased desire to respond. Study 2 compared real-world cross-ideological interactions to homogeneous interactions on Twitter. When talking to people with the same ideology, receiving high numbers of retweets and likes had little impact on outrage. Unsurprisingly, seeing the ideas of people who share our worldview spread does not predict angry replies. However, when replying to users with a different ideology, greater virality predicted angrier and more negative moral-emotional replies. Furthermore, results suggested that the difference between cross-cutting and homogeneous interactions was more pronounced in replies to especially influential twitter users. Lastly, Study 3 demonstrated that people strategically use outrage when they consciously try to coordinate audiences against someone else (relative to when they try to receive upvotes themselves). Results also suggested anonymity may increase the strategic use of outrage slightly further. Combined these results demonstrate that outrage is triggered by the threat of viral, opposing views and used consciously as a strategic tool to inhibit their spread.

The Roots of Digital Outrage

These results suggest that the information social media communicate about the spread of ideas may be one driver of digital outrage. I theorize that quantified, public markers of social rewards, such as likes and retweets, contain valuable information about social consensus. Outrage is one of our most valuable tools for coordinating and building consensus around our side in disagreements. When social media bombards us with information about the spread of opposing values, we feel and express outrage to secure the moral high ground.

Virtue signaling suggests an alternative explanation for the present results. In Study 2, it is possible that Twitter users replied with more outrage to viral tweets because they felt they provided the best opportunities to signal personal reputation. In other words, perhaps Tweets that are getting more attention seem like the best opportunity to signal personal virtues through outrage. Similarly in Study 1, perhaps the effects of virality upon subjective outrage reflected people following a heuristic that responding with outrage to viral content is a good strategy for signaling reputation. This alternative explanation is difficult to rule out. In fact, the argument that outrage is always, to some degree, driven by an implicit goal to make oneself look better borders on unfalsifiable. However, it is unlikely that the prevalence of outrage expression reduces to one theoretical explanation. People condemn and blame others for a variety of reasons. The desire to signal personal reputation undoubtedly motivates some of the outrage that takes place on social media. However, witnessing opposing content go viral increases the motivation to act; claiming that those motivations are always implicitly linked to self-promotion is reductive and ignores other perspective on condemnation that offer more straightforward explanations. Condemnation accomplishes more than just signaling virtue. It impacts the sides people choose in conflict, who holds political power, and which policies eventually become reality.

Implications for Designing Digital Communities

In 2019, Instagram began experimenting with eliminating the like count from their platform—allowing users to see their own, but the not the tally of likes others receive. Adam Mosseri, CEO of Instagram, indicated the change was "about creating a less pressured environment" (Meisenzahl, 2019). The present results may lend additional support to Mosseri's rationale. Not only do likes signal information about how people view one's personal posts, they communicate how the ideas of our opponents spread and gain influence. In political contexts this may subjectively raise the stakes of cross-cutting conversations. It is not just personal reputation that is on the line. When beliefs we see as dangerous go on to gain support and spread virally, we may feel like stopping their spread partly depends upon what we say in response. The present results demonstrate that when consciously trying to coordinate audiences against a target, we become increasingly outraged.

Masking like and retweet buttons seem like a natural solution to these problems, but design considerations should holistically consider their social psychological effects. The like button also has many positive qualities. It allows us to show our support for friends who need it, to voice our opinion and democratically afford power to the movements we support, and to filter information by topics that are having the most social impact. Design choices should consider all of the ways like buttons impact how we interact with and consume information. However, for interactions specifically involving people who disagree politically, the negative impacts of like buttons may more clearly outweigh the positives. Understanding people across the political divide is already an incredibly difficult task for most people. Keeping score of who is "winning" the conversation is unlikely to make cross-cutting conversations any easier. Thus digital

environments hoping to foster productive and diverse political conversations should consider how likes and upvotes may inhibit civility.

Limitations and Future Directions

The present studies provide evidence from a variety of samples (university students, Mechanical Turk worker and Twitter users) and contexts (online surveys and real-world social media platforms). However, they are not without limitations. First, the effect size of virality upon outrage was small in both studies 1 and 2. It is possible, however, that the effect of virality does not primarily manifest as *subjectively felt* outrage (as measured in Study 1). Instead, virality may have a larger impact upon the *expression of outrage*. Study 3, for example, observed increased outrage expression in the absence of differences in subjective outrage. This may suggest that outrage expression is performative and strategic. Alternatively, people may be reluctant to admit that offensive content has upset them, limiting the effectiveness of self-report outrage measures. Small effect sizes were also observed in Study 2. However, lexicon-based methods of text analyses (such as those used in Study 2) are known to have more error than other methods of sentiment analyses, such as machine learning classifiers (Hailong & Wenyan, 2018). Although they also have advantages such as not requiring human labeled training data which is subject to bias. In short, text analysis of moral language is still in its infancy, and errors in measurement make estimating effect sizes precisely difficult.

Study 1 failed to find the predicted attenuating effect of downvotes upon outrage, despite decreasing the perceived impact of the offensive tweets. Since Twitter does not have a downvote button participants had to imagine that one exists. The hypothetical nature of the downvotes suggests the present results may underestimate the impact of actual downvotes. Thus the extent that downvotes mitigate the threat of virally spreading ideas is not entirely clear. Future studies

would benefit from more realistic manipulations or comparisons across existing social media platforms that contain (or do not contain) public markers of social disapproval.

The measure of perceived impact in Study 1 also did not consistently relate to outrage or desire to act. Collapsing all ratings of perceived impact and desire to act for each participant, perceived impact shared a small, positive relationship with desire to act, $r(238) = .16$, $p = .01$, and no relationship with subjective outrage $r(238) = .03$, $p = .62$. This small relationship could stem from an issue with how perceived impact was measured. For example, people who find the tweet most outrage inducing might also be the most reluctant to admit that the tweet has a large amount of support. Alternatively, the effect of virality on reactions could simply be small. An effect size of $r = .16$ is not completely out of line with the size of the observed effect of the virality manipulation upon desire to act.

While the present studies examined the impact of *shifts* in public markers of approval upon outrage, they did not manipulate the *mere presence* of "likes" or "retweets." In other words, the existence of markers of social approval may trigger more competitive and less civil mindsets on its own, even without others' comments going viral. Future studies could both manipulate the presence of "like" and "share" buttons and compare cross-cutting conversations on existing social media sites that contain or lack them.

Study 2 only examined conversations surrounding one political topic. It is possible that people use different types of language depending on the issue being discussed. Thus effects of virality upon moral language should be replicated in different political topics. Lastly, Study 1 found suggestive evidence of different effects for liberal versus conservative participants. However, the sample size of conservative participants was small, and the materials used for conservatives differed from those used for liberals (though they were perceived as equally

ideologically extreme). This makes it difficult to draw conclusions about whether the effect of virality is ideologically symmetrical. Future research should further explore ideological asymmetry in the effects of virality upon outrage.

Conclusions

Slate magazine labelled 2014 "The Year of Outrage." At the time, mobs of outraged tweeters calling for the firing or boycotting of controversial figures felt abnormal. But every year since 2014 is likely just as deserving of *Slate's* award. Conversations about hot-button issues increasingly take place through a computer screen, and the features present in online interactions have changed considerably since message boards like *usenet* reigned supreme. Now when we discuss topics tied to our core ideas about right and wrong, everything we say and share is accompanied by "points" telling us, and whoever else is watching, exactly what the masses think about our side of a dispute. This raises the stakes of conversations about sensitive topics and may cause us to use tools that inhibit finding common ground, such as outrage and condemnation, to ensure our side comes out on top. As the political climate in the United States grows increasingly hostile and polarized (PEW, 2016), we must understand how the contexts of our conversations may further impede their productivity. Restricting features that make conversations feel more like competitions may be one effective strategy for improve understanding across the ideological divide.

REFERENCES

Aiello, L. M., Barrat, A., Schifanella, R., Cattuto, C., Markines, B., & Menczer, F. (2012). Friendship Prediction and Homophily in Social Media. *ACM Trans. Web*, *6*(2), 9:1–9:33. https://doi.org/10.1145/2180861.2180866

Alhabash, S., & McAlister, A. R. (2015). Redefining virality in less broad strokes: Predicting viral behavioral intentions from motivations and uses of Facebook and Twitter. *New Media & Society*, *17*(8), 1317–1339. https://doi.org/10.1177/1461444814523726

Allison, S. T., Messick, D. M., & Goethals, G. R. (1989). On Being Better but not Smarter than Others: The Muhammad Ali Effect. *Social Cognition*, *7*(3), 275–295. https://doi.org/10.1521/soco.1989.7.3.275

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, *211*(4489), 1390–1396. https://doi.org/10.1126/science.7466396

Badawy, A., Ferrara, E., & Lerman, K. (2018). Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 258–265. https://doi.org/10.1109/ASONAM.2018.8508646

Bae, S. Y. (2014). *Examining Political Engagement in an Age of Social Media*.

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130-1132. doi:10.1126/science.aaa1160

Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, *23*(1), 76-91.

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior*, *27*(5), 325–344. https://doi.org/10.1016/j.evolhumbehav.2006.01.003

Barclay, P. (2013). Strategies for cooperation in biological markets, especially for humans. *Evolution and Human Behavior*, *34*(3), 164–175. https://doi.org/10.1016/j.evolhumbehav.2013.02.002

Barclay, P. (2016). Biological markets and the effects of partner choice on cooperation and friendship. *Current Opinion in Psychology*, *7*, 33–38. https://doi.org/10.1016/j.copsyc.2015.07.012

Barclay, P., & Kiyonari, T. (2014). Why sanction? Functional causes of punishment and reward. *Reward and Punishment in Social Dilemmas*, 182-196. doi:10.1093/acprof:oso/9780199300730.003.0010

Batson, C. D., & Thompson, E. R. (2001). Why Don't Moral People Act Morally? Motivational Considerations. *Current Directions in Psychological Science*, *10*(2), 54–57. https://doi.org/10.1111/1467-8721.00114

Baumard, N., André, J. B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences*, *36*(1), 59–78. https://doi.org/10.1017/S0140525X11002202

Baumeister, R. F. (1982). A self-presentational view of social phenomena. *Psychological Bulletin,91*(1), 3-26. doi:10.1037//0033-2909.91.1.3

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313-7318. doi:10.1073/pnas.1618923114

Brady, W. J., Crockett, M., & Van Bavel, J. J. (2019). *The MAD Model of Moral Contagion: The role of motivation, attention and design in the spread of moralized content online*. Manuscript Submitted for Publication. https://doi.org/10.31234/osf.io/pz9g6

Bernstein, L. (1992). Opting out of the Legal System: Extralegal Contractual Relations in the Diamond Industry. *The Journal of Legal Studies*, *21*(1), 115–157. https://doi.org/10.1086/467902

Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., … Dredze, M. (2018). Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate. *American Journal of Public Health*, *108*(10), 1378–1384. https://doi.org/10.2105/AJPH.2018.304567

Brosig, J. (2002). Identifying cooperative behavior: Some experimental results in a prisoner's dilemma game. *Journal of Economic Behavior & Organization*, *47*(3), 275–290. https://doi.org/10.1016/S0167-2681(01)00211-6

Bocian, K., & Wojciszke, B. (2014). Self-Interest Bias in Moral Judgments of Others' Actions. *Personality and Social Psychology Bulletin*, *40*(7), 898–909. https://doi.org/10.1177/0146167214529800

Burrow, A. L., & Rainone, N. (2017). How many likes did I get?: Purpose moderates links between positive social media feedback and self-esteem. *Journal of Experimental Social Psychology*, *69*, 232–236. https://doi.org/10.1016/j.jesp.2016.09.005

Corneille, O., Yzerbyt, V. Y., Rogier, A., & Buidin, G. (2001). Threat and the group attribution error: When threat elicits judgments of extremity and homogeneity. *Personality and Social Psychology Bulletin, 27*(4), 437-446. doi:10.1177/0146167201274005

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour, 1*(11), 769-771. doi:10.1038/s41562-017-0213-3

Curtis, V., & Biran, A. (2001). Dirt, disgust, and disease: Is hygiene in our genes? *Perspectives in Biology and Medicine, 44*(1), 17-31. doi:10.1353/pbm.2001.0001

Dickinson, A., Nicholas, D. J., & Adams, C. D. (1983). The effect of the instrumental training contingency on susceptibility to reinforcer devaluation. *The Quarterly Journal of Experimental Psychology Section B, 35*(1), 35–51. https://doi.org/10.1080/14640748308400912

Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management Science, 32*(5), 554-571. doi:10.1287/mnsc.32.5.554

Descioli, P. (2016). The side-taking hypothesis for moral judgment. *Current Opinion in Psychology, 7*, 23-27. doi:10.1016/j.copsyc.2015.07.002

Descioli, P., & Kurzban, R. (2009). Mysteries of morality. *Cognition, 112*(2), 281-299. doi:10.1016/j.cognition.2009.05.008

Descioli, P., & Kurzban, R. (2013). A solution to the mysteries of morality. *Psychological Bulletin, 139*(2), 477-496. doi:10.1037/a0029065

DeScioli, P., Massenkoff, M., Shaw, A., Petersen, M. B., & Kurzban, R. (2014). Equity or equality? Moral judgments follow the money. *Proceedings of the Royal Society B: Biological Sciences, 281*(1797), 20142112. https://doi.org/10.1098/rspb.2014.2112

Eady, G., Nagler, J., Guess, A., Zilinsky, J., & Tucker, J. A. (2019). How Many People Live in Political Bubbles on Social Media? Evidence From Linked Survey and Twitter Data. *SAGE Open*, *9*(1), 2158244019832705. https://doi.org/10.1177/2158244019832705

Everett, J. A. C., Faber, N. S., Savulescu, J., & Crockett, M. J. (2018). The costs of being consequentialist: Social inference from instrumental harm and impartial beneficence. *Journal of Experimental Social Psychology*, *79*, 200–216. https://doi.org/10.1016/j.jesp.2018.07.004

Everett, J. A. C., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General*, *145*(6), 772–787. https://doi.org/10.1037/xge0000165

Fiske, A. P., & Rai, T. S. (2015). *Virtuous violence: Hurting and killing to create, sustain, end, and honor social relationships*. Cambridge: Cambridge University Press.

Frimer, J. A., Skitka, L. J., & Motyl, M. (2017). Liberals and conservatives are similarly motivated to avoid exposure to one another's opinions. *Journal of Experimental Social Psychology*, *72*, 1–12. https://doi.org/10.1016/j.jesp.2017.04.003

Fu, F., Hauert, C., Nowak, M. A., & Wang, L. (2008). Reputation-based partner choice promotes cooperation in social networks. *Physical Review E*, *78*(2), 026117. https://doi.org/10.1103/PhysRevE.78.026117

Gearhart, S., & Zhang, W. (2014). Gay Bullying and Online Opinion Expression: Testing Spiral of Silence in the Social Media Environment. *Social Science Computer Review*, *32*(1), 18–36. https://doi.org/10.1177/0894439313504261

Grubbs, J. B., Warmke, B., Tosi, J., James, A. S., & Campbell, W. K. (2019). *Moral Grandstanding in Public Discourse: Status-seeking Motives as a Potential Explanatory*

Mechanism in Predicting Conflict. Manuscript Submitted for Publication. https://doi.org/10.31234/osf.io/gnaj5

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96*(5), 1029-1046. doi:10.1037/a0015141

Goldenberg, A., Gross, J., & Garcia, D. (2018). Emotional Sharing on Social Media: How Twitter Replies Contribute to Increased Emotional Intensity. Grafen, A. (1984). Natural selection, kin selection and group selection. *Behavioural ecology: An evolutionary approach, 2*, 62-84.

Guala, F. (2012). Strong reciprocity is real, but there is no evidence that uncoordinated costly punishment sustains cooperation in the wild. *Behavioral and Brain Sciences, 35*(01), 45-59. doi:10.1017/s0140525x1100166x

Haidt, J., Rosenberg, E., & Hom, H. (2003). Differentiating diversities: moral diversity is not like other kinds. *Journal of Applied Social Psychology, 33*(1), 1-36. doi:10.1111/j.1559-1816.2003.tb02071.x

Halberstam, Y., & Knight, B. (2014). Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter. doi:10.3386/w20681

Hampton, K., Rainie, L., Lu, W., Dwyer, M., Shin, I., & Purcell, K. (2014, August 26). Social Media and the 'Spiral of Silence.' Retrieved October 4, 2019, from Pew Research Center: Internet, Science & Tech website: https://www.pewinternet.org/2014/08/26/social-media-and-the-spiral-of-silence/

Hayat, T., & Samuel-Azran, T. (2017). "You too, second screeners?" second screeners' echo chambers during the 2016 U.S. elections primaries. *Journal of Broadcasting & Electronic Media, 61*(2), 291-308. doi:10.1080/08838151.2017.1309417

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial Punishment Across Societies. *Science, 319*(5868), 1362–1367. https://doi.org/10.1126/science.1153808

Hiltz, S. R., Johnson, K., & Turoff, M. (1986). Experiments in group decision making communication process and outcome in face-to-face versus computerized conferences. *Human Communication Research, 13*(2), 225-252. doi:10.1111/j.1468-2958.1986.tb00104.x

Himelboim, I. (2014). Political television hosts on Twitter: examining patterns of interconnectivity and self-exposure in Twitter political talk networks. *Journal of Broadcasting & Electronic Media, 58*(1), 76-96. doi:10.1080/08838151.2013.875017

Himelboim, I., Mccreery, S., & Smith, M. (2013). Birds of a feather Tweet together: Integrating network and content analyses to examine cross-ideology exposure on Twitter. *Journal of Computer-Mediated Communication, 18*(2), 40-60. doi:10.1111/jcc4.12001

Hofmann, W., Skitka, L., Wisneski, D., & Brandt, M. (2016). Morality in everyday life. *Science*. doi:10.1037/e512142015-287

Hutcherson, C. A., & Gross, J. J. (2011). The moral emotions: A social–functionalist account of anger, disgust, and contempt. *Journal of Personality and Social Psychology, 100*(4), 719-737. doi:10.1037/a0022408

Iyengar, S., & Hahn, K. S. (2009). Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use. *Journal of Communication, 59*(1), 19–39. https://doi.org/10.1111/j.1460-2466.2008.01402.x

Jacobs, D. (1978). Inequality and the legal order: An Ecological test of the conflict model. *Social Problems, 25*(5), 515-525. doi:10.1525/sp.1978.25.5.03a00060

Johnen, M., Jungblut, M., & Ziegele, M. (2018). The digital outcry: What incites participation behavior in an online firestorm? *New Media & Society, 20*(9), 3140–3160. https://doi.org/10.1177/1461444817741883

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature, 530*(7591), 473-476. doi:10.1038/nature16981

Jordan, J. J., & Rand, D. G. (2017). The drive to appear trustworthy shapes punishment and moral outrage in one-shot anonymous interactions. *SSRN Electronic Journal*. doi:10.2139/ssrn.2969063

Katzir, M. (2017, January). *Disgust as an essentialist emotion signals non-violent out-grouping with potentially low social costs.* Poster presented at the Society for Personality and Social Psychology Annual Conference. San Antonio, TX.

Kirkman, B. L., & Mathieu, J. E. (2005). The dimensions and antecedents of team virtuality. *Journal of Management, 31*(5), 700-718. doi:10.1177/0149206305279113

Konishi, N., Oe, T., Shimizu, H., Tanaka, K., & Ohtsubo, Y. (2017). Perceived shared condemnation intensifies punitive moral emotions. *Scientific Reports, 7*(1). doi:10.1038/s41598-017-07916-z

Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior, 28*(2), 75–84. https://doi.org/10.1016/j.evolhumbehav.2006.06.001

Kurzban, R., Dukes, A., & Weeden, J. (2010). Sex, drugs and moral goals: Reproductive strategies and views about recreational drugs. *Proceedings of the Royal Society B: Biological Sciences, 277*(1699), 3501–3508. https://doi.org/10.1098/rspb.2010.0608

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? *Proceedings of the 19th International Conference on World Wide Web - WWW '10*, 591. https://doi.org/10.1145/1772690.1772751

Lee, J. K., Choi, J., Kim, C., & Kim, Y. (2014). Social Media, Network Heterogeneity, and Opinion Polarization. *Journal of Communication, 64*(4), 702–722. https://doi.org/10.1111/jcom.12077

Lea, M., Spears, R., & de Groot, D. D. (2001). Knowing me, knowing you: Anonymity effects on social identity processes within groups. *Personality and Social Psychology Bulletin, 27*(5), 526-537. doi:10.1177/0146167201275002

Leach, C. W., Ellemers, N., & Barreto, M. (2007). Group virtue: The importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of Personality and Social Psychology, 93*(2), 234-249. doi:10.1037/0022-3514.93.2.234

Leary, M. R., & Kowalski, R. M. (1990). Impression management: A literature review and two-component model. *Psychological Bulletin, 107*(1), 34-47. doi:10.1037//0033-2909.107.1.34

Liang, H. (2014). The organizational principles of online political discussion: A relational event stream model for analysis of web forum deliberation. *Human Communication Research, 40*(4), 483-507. doi:10.1111/hcre.12034

McAdams, R. H. (1997). The Origin, Development, and Regulation of Norms. *Michigan Law Review, 96*, 338.

McKenna, K. Y., & Bargh, J. A. (2002). *Consequences of the Internet for self and society: Is social life being transformed?* Oxford: Blackwell.

Meshi, D., Morawetz, C., & Heekeren, H. R. (2013). Nucleus accumbens response to gains in reputation for the self relative to gains for others predicts social media use. *Frontiers in Human Neuroscience, 7*. https://doi.org/10.3389/fnhum.2013.00439

McKenna, K. Y. A., Green, A. S., & Gleason, M. E. J. (2002). Relationship Formation on the Internet: What's the Big Attraction? *Journal of Social Issues*, *58*(1), 9–31. https://doi.org/10.1111/1540-4560.00246

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences, 11*(4), 143-152. doi:10.1016/j.tics.2006.12.007

Molho, C., Tybur, J. M., Güler, E., Balliet, D., & Hofmann, W. (2017). Disgust and anger relate to different aggressive responses to moral violations. *Psychological Science*, *28*(5), 609-619. doi:10.1177/0956797617692000

Motyl, M., Iyer, R., Oishi, S., Trawalter, S., & Nosek, B. A. (2014). How ideological migration geographically segregates groups. *Journal of Experimental Social Psychology*, *51*, 1-14. doi:10.1016/j.jesp.2013.10.010

Munson, S. A., & Resnick, P. (2010). Presenting Diverse Political Opinions: How and How Much. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1457–1466. https://doi.org/10.1145/1753326.1753543

Nelissen, R. M. A. (2008). The price you pay: Cost-dependent reputation effects of altruistic punishment. *Evolution and Human Behavior*, *29*(4), 242–248. https://doi.org/10.1016/j.evolhumbehav.2008.01.001

Newman, G. E., Bloom, P., & Knobe, J. (2014). Value Judgments and the True Self. *Personality and Social Psychology Bulletin*, *40*(2), 203–216. https://doi.org/10.1177/0146167213508791

Noë, R., & Hammerstein, P. (1994). Biological markets: Supply and demand determine the effect of partner choice in cooperation, mutualism and mating. *Behavioral Ecology and Sociobiology*, *35*(1), 1–11. https://doi.org/10.1007/BF00167053

Omernick, E., & Sood, S. O. (2013). The impact of anonymity in online communities. *2013 International Conference on Social Computing*. doi:10.1109/socialcom.2013.80

Opotow, S. (1990). Moral exclusion and injustice: An introduction. *Journal of Social Issues*, *46*(1), 1-20. doi:10.1111/j.1540-4560.1990.tb00268.x

Ostrom, T. M., & Sedikides, C. (1992). Out-group homogeneity effects in natural and minimal groups. *Psychological Bulletin, 112*(3), 536-552. doi:10.1037//0033-2909.112.3.536

Pariser, E. (2011). *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC2015*. Retrieved from https://repositories.lib.utexas.edu/handle/2152/31333

Petersen, M. B. (2013). Moralization as protection against exploitation: Do individuals without allies moralize more? *Evolution and Human Behavior*, *34*(2), 78-85. doi:10.1016/j.evolhumbehav.2012.09.006

Pew (2016). *Partisanship and political animosity in 2016.* Retrieved from http://www.people-press.org/2016/06/22/partisanship-and-political-animosity-in-2016/

Pew (2018). *Activism in the social media age.* Retrieved from https://www.pewinternet.org/2018/07/11/public-attitudes-toward-political-engagement-on-social-media/

Pfeffer, J., Zorbach, T., & Carley, K. M. (2014). Understanding online firestorms: Negative word-of-mouth dynamics in social media networks. *Journal of Marketing Communications*, *20*(1-2), 117-128.

Piazza, J., & Bering, J. M. (2008). Concerns about reputation via gossip promote generous allocations in an economic game. *Evolution and Human Behavior*, *29*(3), 172-178. doi:10.1016/j.evolhumbehav.2007.12.002

Pinker, S. (2010). *The better angels of our nature: Why violence has declined*. New York: Penguin.

Postmes, T., Spears, R., & Lea, M. (2002). Intergroup differentiation in computer-mediated communication: Effects of depersonalization. *Group Dynamics: Theory, Research, and Practice*, *6*(1), 3-16. doi:10.1037//1089-2699.6.1.3

Pratto, F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology*, *67*(4), 741-763. doi:10.1037//0022-3514.67.4.741

Quintelier, K. J. P., Ishii, K., Weeden, J., Kurzban, R., & Braeckman, J. (2013). Individual Differences in Reproductive Strategy are Related to Views about Recreational Drug Use in Belgium, The Netherlands, and Japan. *Human Nature*, *24*(2), 196–217. https://doi.org/10.1007/s12110-013-9165-0

Raihani, N. J., & Bshary, R. (2015). The reputation of punishers. *Trends in Ecology & Evolution*, *30*(2), 98-103. doi:10.1016/j.tree.2014.12.003

Rains, S. A., Kenski, K., Coe, K., & Harwood, J. (2017). Incivility and political identity on the Internet: Intergroup factors as predictors of incivility in discussions of news online. *Journal of Computer-Mediated Communication, 22*(4), 163-178. doi:10.1111/jcc4.12191

Rand, D. G., & Nowak, M. A. (2011). The evolution of antisocial punishment in optional public goods games. *Nature Communications, 2*, 434. doi:10.1038/ncomms1442

Reicher, S. D., Spears, R., & Postmes, T. (1995). A social identity model of deindividuation phenomena. *European Review of Social Psychology, 6*(1), 161-198. doi:10.1080/14792779443000049

Rockenbach, B., & Milinski, M. (2011). To qualify as a social partner, humans hide severe punishment, although their observed cooperativeness is decisive. *Proceedings of the National Academy of Sciences*, *108*(45), 18307–18312. https://doi.org/10.1073/pnas.1108996108

Rom, S. C., & Conway, P. (2018). The strategic moral self: Self-presentation shapes moral dilemma judgments. *Journal of Experimental Social Psychology*, *74*, 24–37. https://doi.org/10.1016/j.jesp.2017.08.003

Rothgerber, H. (1997). External intergroup threat as an antecedent to perceptions in in-group and out-group homogeneity. *Journal of Personality and Social Psychology, 73*(6), 1206-1212. doi:10.1037//0022-3514.73.6.1206

Ryan, T. J. (2014). Reconsidering moral issues in politics. *The Journal of Politics, 76*(2), 380-397. doi:10.1017/s0022381613001357

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience*, *15*(8), 549–562. https://doi.org/10.1038/nrn3776

Santos, M. D., Rankin, D. J., & Wedekind, C. (2013). Human cooperation based on punishment reputation. *Evolution*, *67*(8), 2446-2450. doi:10.1111/evo.12108

Sarkissian, H., Park, J., Tien, D., Wright, J. C., & Knobe, J. (2011). Folk Moral Relativism. *Mind & Language*, *26*(4), 482–505. https://doi.org/10.1111/j.1468-0017.2011.01428.x

Sawaoka, T., & Monin, B. (2018). The Paradox of Viral Outrage. *Psychological Science*, *29*(10), 1665–1678. https://doi.org/10.1177/0956797618780658

Schlenker, B. (1992). Interpersonal processes involving impression regulation and management. *Annual Review of Psychology*, *43*(1), 133-168. doi:10.1146/annurev.psych.43.1.133

Sherman, L. E., Payton, A. A., Hernandez, L. M., Greenfield, P. M., & Dapretto, M. (2016). The Power of the Like in Adolescence: Effects of Peer Influence on Neural and Behavioral Responses to Social Media. *Psychological Science*, *27*(7), 1027–1035. https://doi.org/10.1177/0956797616645673

Short, J., Christie, B., & Williams, E. (1976). *The social psychology of telecommunications*. London: Wiley.

Shugars, S., & Beauchamp, N. (2019). Why Keep Arguing? Predicting Engagement in Political Conversations Online. *SAGE Open*, *9*(1), 215824401982885. https://doi.org/10.1177/2158244019828850

Simon, B. (1992). The perception of ingroup and outgroup homogeneity: Reintroducing the intergroup context. *European Review of Social Psychology*, *3*(1), 1-30. doi:10.1080/14792779243000005

Skitka, L. J. (2010). The psychology of moral conviction. *Social and Personality Psychology Compass*. 4(4), 267–281. doi: 10.1111/j.1751-9004.2010.00254.x

Skitka, L. J., Bauman, C. W., & Sargis, E. G. (2005). Moral conviction: Another contributor to attitude strength or something more? *Journal of Personality and Social Psychology, 88*(6), 895-917. doi:10.1037/0022-3514.88.6.895

Springer, N., Engelmann, I., & Pfaffinger, C. (2015). User comments: motives and inhibitors to write and read. *Information, Communication & Society, 18*(7), 798-815. doi:10.1080/1369118x.2014.997268

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition*, *131*(1), 159-171. doi: 10.1037/e505052014-054

Strohminger, N., & Nichols, S. (2015). Neurodegeneration and identity. *Psychological Science, 26*(9), 1469-1479. doi:10.1177/0956797615592381

Suler, J. (2004). The online disinhibition effect. *CyberPsychology & Behavior, 7*(3), 321-326. doi:10.1089/1094931041291295

Sullivan, D., Landau, M. J., Branscombe, N. R., & Rothschild, Z. K. (2012). Competitive victimhood as a response to accusations of ingroup harm doing. *Journal of Personality and Social Psychology*, *102*(4), 778–795. https://doi.org/10.1037/a0026573

Sunstein, C. R. (2018). *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press.

Sylwester, K., & Roberts, G. (2013). Reputation-based partner choice is an effective alternative to indirect reciprocity in solving social dilemmas. *Evolution and Human Behavior*, *34*(3), 201–206. https://doi.org/10.1016/j.evolhumbehav.2012.11.009

Tetlock, P. E. (2003). Thinking the unthinkable: Sacred values and taboo cognitions. *Trends in Cognitive Sciences, 7*(7), 320-324. doi:10.1016/s1364-6613(03)00135-9

Tooby, J., & Cosmides, L. (2010). Groups in mind: The coalitional roots of war and morality. *Human Morality and Sociality,* 191-234. doi:10.1007/978-1-137-05001-4_8

Tucker, J. A., Guess, A., Barbera, P., Vaccari, C., Siegel, A., Sanovich, S., … Nyhan, B. (2018). *Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature* (SSRN Scholarly Paper No. ID 3144139). Retrieved from Social Science Research Network website: https://papers.ssrn.com/abstract=3144139

Tybur, J. M., Lieberman, D., & Griskevicius, V. (2009). Microbes, mating, and morality: Individual differences in three functional domains of disgust. *Journal of Personality and Social Psychology*, *97*(1), 103-122. doi:10.1037/a0015474

Uhlmann, E. L., Zhu, L. (Lei), & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition*, *126*(2), 326–334. https://doi.org/10.1016/j.cognition.2012.10.005

Vandeselaere, N. (1991). The impact of in-Group and out-Group homogeneity/heterogeneity upon intergroup relations. *Basic and Applied Social Psychology*, *12*(3), 291-301. doi:10.1207/s15324834basp1203_4

Verplaetse, J., Vanneste, S., & Braeckman, J. (2007). You can judge a book by its cover: The sequel.: A kernel of truth in predictive cheating detection. *Evolution and Human Behavior*, *28*(4), 260–271. https://doi.org/10.1016/j.evolhumbehav.2007.04.006

Vonasch, A. J., Reynolds, T., Winegard, B. M., & Baumeister, R. F. (2017). Death before dishonor. *Social Psychological and Personality Science,* doi:10.1177/1948550617720271

Wallace, P. (1999). *The Psychology of the Internet*. Cambridge University Press.

Weeden, J. (2003). *Genetic interests, life histories, and *attitudes towards abortion*. Retrieved from https://repository.upenn.edu/dissertations/AAI3087480

Wilder, D. A. (1986). Social categorization: implications for creation and reduction of intergroup bias. *Advances in Experimental Social Psychology*, 291-355. doi:10.1016/s0065-2601(08)60217-8

Yardi, S., & Boyd, D. (2010). Dynamic debates: An analysis of group polarization over time on Twitter. *Bulletin of Science, Technology & Society*, *30*(5), 316-327. doi:10.1177/0270467610380011

Young, I. F., & Sullivan, D. (2016). Competitive victimhood: A review of the theoretical and empirical literature. *Current Opinion in Psychology*, *11*, 30-34. doi:10.1016/j.copsyc.2016.04.004

APPENDICES

Appendix A: Study 1 Prescreen

When it comes to politics in general, how would you describe yourself?

- Very liberal
- Liberal
- Somewhat liberal
- Middle of the Road
- Somewhat conservative
- Conservative
- Very conservative

In general, do you lean more Democrat or Republican?

- Democrat
- Republican

For which of the following social media platforms do you have an account (check all that apply)?

- Twitter
- Facebook
- Snapchat
- Instagram

When did you most recently check Twitter?

- Never
- More than a month ago
- More than a week ago
- More than a day ago
- Today

Are you of Hispanic or Latino or Spanish Origin?

- Yes
- No

What is your race?

- American Indian or Alaska Native
- Asian
- Black or African American
- Native Hawaiian or Other Pacific Islander
- White
- Other

Appendix B: Study 1 Text for Outrage Inducing Tweets

Conservative (presented to liberal participants)

"Murder and assault are spiraling out of control in cities all over the country thanks to 3rd world illegals who shouldn't even be here"

"Sick and tired of seeing gays touching and kissing anytime I turn on the TV. Makes me sick"

"Mothers who murder their unborn babies to avoid the consequences of their actions should be locked away in prison. The one's who do it over and over (serial killers) should get the chair"

"All these HS kids whining about guns. Psychopaths shooting up schools have nothing to do with my 2nd amendment rights."

Liberal (presented to conservative participants)

"Open borders, abolish ICE, citizenship for illegals, yes to all of it. Anything that stops this country from being run by a bunch of old white men"

"Until the baby is crying outside of the womb, every woman should be able to terminate her pregnancy for whatever reason NO QUESTIONS ASKED"

"Awe poor police officers crying over a taste of their medicine. Next time I hope protesters bring more than eggs to throw at them"

"No one has any business owning a gun. For. Any. Reason. They were made for one purpose: killing innocent creatures."

Appendix C: Study 1 Tweet Manipulations Examples

Example of high favorite/retweet accumulation; No visible backlash

Hour 0

Hour 3

Hour 6

Hour 9

Hour 12

Example of high favorite/retweet accumulation; Downvote backlash

Example of high favorite/retweet accumulation; Comment backlash

Hour 0

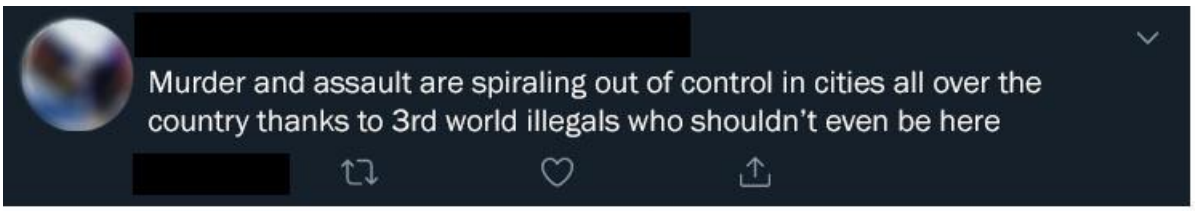
Hour 3

Hour 6

Hour 9

Hour 12

Example of low favorite/retweet accumulation; No backlash



Hour 0

Hour 3

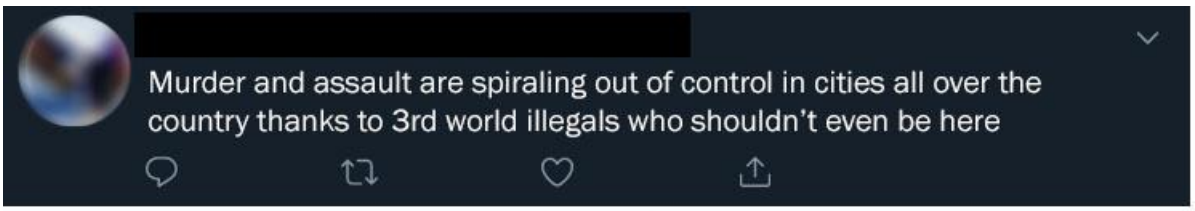
Hour 6

Hour 9

Hour 12

Example of low favorite/retweet accumulation; Downvote backlash

Example of low favorite/retweet accumulation; Comment backlash



Hour 0

Hour 3

Hour 6

Hour 9

Hour 12

Appendix D: Self-reported outcomes and manipulation checks

Study 1 prompt [Study 3 prompt]:

To what extent did what you saw on the previous page make you feel…

[To what extent does the above comment make you feel…]

Not at all angry	1	2	3	4	5	6	7 *Very angry*
Not at all offended	1	2	3	4	5	6	7 *Very offended*
Not at all outraged	1	2	3	4	5	6	7 *Very outraged*
Not at all upset	1	2	3	4	5	6	7 *Very upset*
Not at all satisfied	1	2	3	4	5	6	7 *Very satisfied*
Not at all pleased	1	2	3	4	5	6	7 *Very pleased*
Not at all afraid	1	2	3	4	5	6	7 *Very afraid*
Not at all threatened	1	2	3	4	5	6	7 *Very threatened*

[Study 1 only]
If you saw this on Twitter would you feel the need to speak up?
Not at all 1 2 3 4 5 6 7 *Very much*

If you saw this on Twitter how likely would you be to write a reply?
Not at all likely 1 2 3 4 5 6 7 *Very likely*

Study 1 Manipulation checks:

1) When it comes to politics, what do you think best describes the person who holds the above account?
   - Very liberal
   - Liberal
   - Somewhat liberal
   - Middle of the Road
   - Somewhat conservative
   - Conservative
   - Very conservative
2) How would you describe the size of this person's Twitter following?
   - Extremely small
   - Small
   - Moderate
   - Large
   - Extremely large
3) To what extent did the tweet gain support over time?
   1 – *Not at all* to 7 – *Very much*
4) To what extent did this tweet influence people?
   1 – *Not at all* to 7 – *Very much*

Appendix E: Moral-Emotional Word Dictionaries

benefit*	Positive Moral-Emotional Words
care	Positive Moral-Emotional Words
caring	Positive Moral-Emotional Words
compassion*	Positive Moral-Emotional Words
devot*	Positive Moral-Emotional Words
faith*	Positive Moral-Emotional Words
good	Positive Moral-Emotional Words
goodness	Positive Moral-Emotional Words
heaven*	Positive Moral-Emotional Words
hero*	Positive Moral-Emotional Words
honest*	Positive Moral-Emotional Words
honor*	Positive Moral-Emotional Words
ideal*	Positive Moral-Emotional Words
loyal*	Positive Moral-Emotional Words
peace*	Positive Moral-Emotional Words
respect	Positive Moral-Emotional Words
safe*	Positive Moral-Emotional Words
save	Positive Moral-Emotional Words
secur*	Positive Moral-Emotional Words
value*	Positive Moral-Emotional Words
virtue*	Positive Moral-Emotional Words
envy*	Negative Moral-Emotional Words
evil*	Negative Moral-Emotional Words
fault*	Negative Moral-Emotional Words
fight*	Negative Moral-Emotional Words
forbid*	Negative Moral-Emotional Words
greed*	Negative Moral-Emotional Words
gross*	Negative Moral-Emotional Words
harm*	Negative Moral-Emotional Words
hate	Negative Moral-Emotional Words
hell	Negative Moral-Emotional Words
hurt*	Negative Moral-Emotional Words
immoral*	Negative Moral-Emotional Words
kill*	Negative Moral-Emotional Words
liar*	Negative Moral-Emotional Words
murder*	Negative Moral-Emotional Words
offend*	Negative Moral-Emotional Words
pain	Negative Moral-Emotional Words
protest	Negative Moral-Emotional Words
punish*	Negative Moral-Emotional Words
rebel*	Negative Moral-Emotional Words
revenge*	Negative Moral-Emotional Words
ruin*	Negative Moral-Emotional Words
shame*	Negative Moral-Emotional Words
sin	Negative Moral-Emotional Words
sinister	Negative Moral-Emotional Words
sins	Negative Moral-Emotional Words
slut*	Negative Moral-Emotional Words
spite*	Negative Moral-Emotional Words
steal*	Negative Moral-Emotional Words
suffer*	Negative Moral-Emotional Words
victim*	Negative Moral-Emotional Words
vile	Negative Moral-Emotional Words
war	Negative Moral-Emotional Words
warring	Negative Moral-Emotional Words
wars	Negative Moral-Emotional Words
whore*	Negative Moral-Emotional Words
wicked*	Negative Moral-Emotional Words
wrong*	Negative Moral-Emotional Words

Appendix F: Study 3 instruction manipulations attention checks

Opponent downvote vs Personal upvote (text altered between conditions is underlined with differences in brackets):

> "We are studying how effective people are at manipulating the number of likes or upvotes they receive on social media. We've collected some comments from other USF students about campus life. We want you to write some replies for us to show participants in a future study, and we will test whether people upvote your comments [your comments make the person you are replying to receive more downvotes]. In other words, on the following pages we want you to write replies that you think will make people upvote your comments [cause people to downvote the person you are replying to]."

Anonymous vs. Identified:

> "Below is an image of how your comment will appear to future participants. USF students participating in our future studies will read the replies you write as depicted below. It is important for you to know that your real name will appear next to the comments that you write, making your identity known to future participants who read whatever replies you write [It is important for you to know that we will collect no data about your identity. Whatever replies you write will be completely anonymous and your identity will be unknown to future participants]."

Attention Check 1:
In your own words, summarize with one sentence the goal of the comments we want you to write on the following pages.

_____

_____

Attention Check 2:
When future participants read the replies you write, what (if anything) will they see next to it?

-your real name
-a profile picture
-your year in school
-there will be no identifying information whatsoever

Appendix G: Screenshots of online message board shown to participants

Practice Comment; Anonymous:

Practice Comment; Identified:

Outrage Inducing Comment; Anonymous:

Outrage Inducing Comment; Identified:

Appendix H: Coding Instructions

First Comment

I like the set up for the gym a lot but oh my god does it get crowded in there. Have to stand around awkwardly for 10 mins just to get a bench somedays

Second Comment

Wow was not expecting for there to be soooo many black people here. Not sure I can get used to this. Weird.

Instructions for rating replies:

How outraged is the person replying?

0 – Not at all outraged.

1 – To a small degree, the comment is critical or condemns the original comment, but does not contain indicators of emotion

2 – Moderate, the person clearly finds the comment offensive or transgressive, but their reply does not contain indicators of strong emotions or affect.

3 – Strong, the person appears to be extremely upset or offended by the content. Reply must contain both indicators of strong emotion ("I feel disgusted right now"; "!?!?!?!?!?!?!!!") and condemnation ("what a racist jerk")

To what extent are they mocking the original comment?

0 – Not at all mocking. Reply is completely serious

1 – To a small degree, not outright making fun of the comment, but contains some levity.

2 – Moderate, makes fun of the comment but also contains some serious elements.

3 – Strong, their entire reply is parodying or making fun of the original comment.

To what extent are they expression moral convictions?

0 – not at all, nothing about their reply seems relevant to their core moral beliefs and convictions

1—to a small degree

2 – A moderate amount

3 – Their comment is explicitly tied to their moral beliefs and convictions

Appendix I: Demographics Used in Study 1 and Study Positive Moral Words

What kind of device did you take the survey on?

- Phone
- Tablet
- Laptop
- Desktop

Type your age (e.g., 21)

____

What is you gender?

- Woman
- Man
- Prefer to self-describe

  _____

[note that the below items about race and ethnicity were included in the pre-screen of Study 1 instead of the end of the survey]

Are you of Hispanic or Latino or Spanish Origin?

- Yes
- No

What is your race?

- American Indian or Alaska Native
- Asian
- Black or African American
- Native Hawaiian or Other Pacific Islander
- White
- Other

Did you find any part of the survey to be confusing or ambiguous?

Appendix J: IRB approval letter

UNIVERSITY OF SOUTH FLORIDA

APPROVAL

November 8, 2019

Curtis Puryear
15501 Bruce B Downs Blvd
Apt 3903
Tampa, FL 33647

Dear Mr. Puryear:

On 11/7/2019, the IRB reviewed and approved the following protocol:

Application Type:	Initial Study
IRB ID:	STUDY000043
Review Type:	Expedited 7
Title:	Digital Interactions
Funding:	None
IND, IDE, or HDE:	None
Approved Protocol and Consent(s)/Assent(s):	• Digital Interactions Protocol v1 10.30.19.docx • Informed Consent v1 10.30.19.pdf; Attached are stamped approved consent documents. Use copies of these documents to document consent.

Within 30 days of the anniversary date of study approval, confirm your research is ongoing by clicking Confirm Ongoing Research in BullsIRB, or if your research is complete, submit a study closure request in BullsIRB by clicking Create Modification/CR.

In conducting this protocol you are required to follow the requirements listed in the INVESTIGATOR MANUAL (HRP-103).

Your study qualifies for a waiver of the requirements for the documentation of informed consent online survey as outlined in the federal regulations at 45 CFR 46.117(c).

Sincerely,

Various Menzel

A PREEMINENT RESEARCH UNIVERSITY

Institutional Review Boards / Research Integrity & Compliance
FWA No. 00001669
University of South Florida / 3702 Spectrum Blvd., Suite 165 / Tampa, FL 33612 / 813-974-5638