aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
1–33

DOI: 10.1177/1745691620917336
www.psychologicalscience.org/PPS
SAGE

The MAD Model of Moral Contagion: The Role of Motivation, Attention, and Design in the Spread of Moralized Content Online

William J. Brady[1], M. J. Crockett[1], and Jay J. Van Bavel[2,3]
[1]Department of Psychology, Yale University; [2]Department of Psychology, New York University; and [3]Center for Neural Science, New York University

Abstract
With more than 3 billion users, online social networks represent an important venue for moral and political discourse and have been used to organize political revolutions, influence elections, and raise awareness of social issues. These examples rely on a common process to be effective: the ability to engage users and spread moralized content through online networks. Here, we review evidence that expressions of moral emotion play an important role in the spread of moralized content (a phenomenon we call *moral contagion*). Next, we propose a psychological model called the motivation, attention, and design (MAD) model to explain moral contagion. The MAD model posits that people have group-identity-based *motivations* to share moral-emotional content, that such content is especially likely to capture our *attention*, and that the *design* of social-media platforms amplifies our natural motivational and cognitive tendencies to spread such content. We review each component of the model (as well as interactions between components) and raise several novel, testable hypotheses that can spark progress on the scientific investigation of civic engagement and activism, political polarization, propaganda and disinformation, and other moralized behaviors in the digital age.

Keywords
morality, emotion, politics, social networks, social media

With more than 3 billion monthly active users, online social networks are an important venue for moral and political discourse across the globe. Social media has been used to organize political revolutions (e.g., the Arab Spring; Lotan et al., 2011), influence presidential elections (e.g., the 2016 U.S. presidential election; Enli, 2017), spread disinformation and political propaganda (Allcott & Gentzkow, 2017; Kollanyi, Howard, & Woolley, 2016; Vosoughi, Roy, & Aral, 2018), and raise awareness of moral issues (Crockett, 2017; Van Der Linden, 2017). Each of these examples ultimately relies on a common process to be effective: the ability to draw user engagement and spread moralized content through online networks. Here, we review recent evidence documenting what type of moralized content is most likely to spread online and then propose a psychological model that helps to explain when, why, and how it spreads.

To understand the spread of "moralized" content, we first offer a working definition of morality. Borrowing from previous work in moral psychology (e.g., Haidt, 2003), we classify content as moralized if it references ideas, objects, or events typically construed in terms of the interests or good of a unit larger than the individual (e.g., society, culture, one's social network). This broad classification allows flexibility in classifying content as moralized regardless of the specifics of the moral content or cultural differences about what is perceived as "right" and "wrong." For example, a social-media message communicating thoughts about gun control in

Corresponding Author:
William J. Brady, Yale University, Department of Psychology, 2 Hillhouse Ave., New Haven, CT 06511
E-mail: william.brady@yale.edu

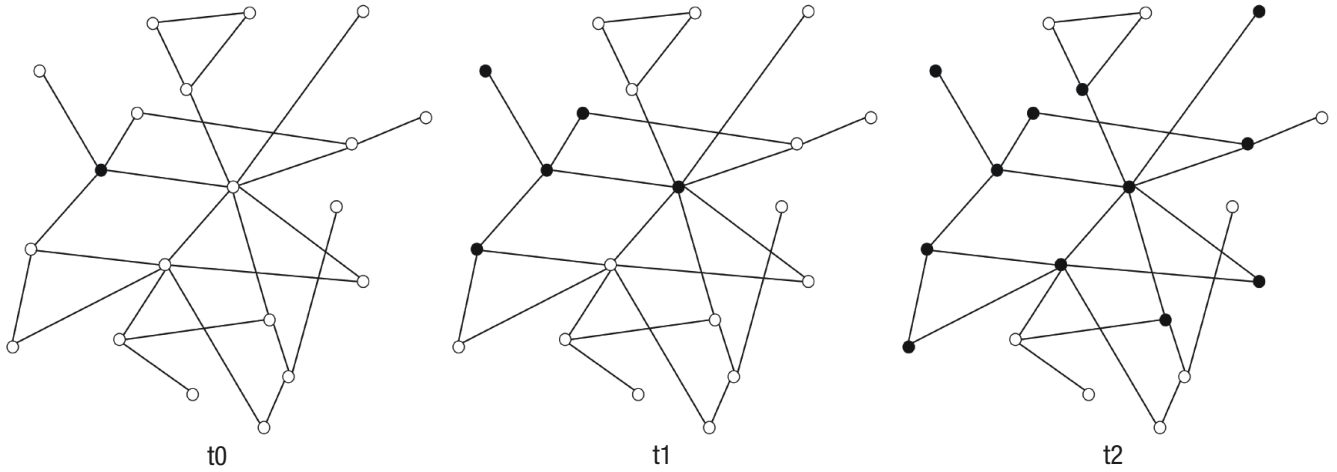

Fig. 1. The rapid spread of information on social media. If one user shares information to their network, it can quickly spread widely on social media. The graph depicts a random network representing social-media users, their friends, and two time points of sharing. Nodes (dots) represent people and ties (lines) represent online relationships; a shaded node indicates that the user has been exposed to the information. If each user shares information to their friends, the information will increase its exposure from one person to the majority of the network in only two rounds of sharing.

America is often construed as moralized content because the topic of gun control is situated in a cultural discussion of whether stricter gun laws are good or bad for American society. On the other hand, a social-media message about cute kittens does not reference a topic that is typically construed in terms of it being good or bad for society. This aspect of construal is crucial to moralization and changes how people evaluate actions (Rozin, 1999; van Bavel, Packer, Haas, & Cunningham, 2012).

Next, we define what counts as "spreading." At the social-network level, the spreading of some phenomena (e.g., information, attitudes, behaviors) is often described as *social contagion* because it resembles the spreading of disease. Each person in a social network can be considered a node that is connected to other nodes through social ties. When one person, or node, becomes "infected" (e.g., they are exposed to a partisan political message), they can easily expose anyone who is socially tied to them (e.g., they share the political message with their social-network friends). This process can rapidly cascade to expose a large portion of the social network to the original content (see Fig. 1). More specifically, social contagion refers to the processes through which attitudes, behaviors, and information spread from one person to another, such as through mimicry (e.g., Chartrand & Bargh, 1999), cognitive appraisal (e.g., Parkinson, 2011), or information diffusion (Bakshy, Rosenn, Marlow, & Adamic, 2012).

For online social networks, the process of social contagion may be particularly important because the process through which content spreads is inherently social and potentially very rapid. Social media allows users to share content they found independently or from people in their social network. People are seven times more likely to share content online when they perceive other people are sharing it (Bakshy et al., 2012), and every share expands the networks that the content can reach. Furthermore, the size of social-media networks is large and transmission is seamless. For instance, if one Twitter user with 1,000 followers shares a message, and if only 10% of those followers share it to their own network of 1,000 followers, the original message will have effortlessly spread to 100,000 users. For these reasons, social media is particularly conducive to the rapid propagation of content. Indeed, the capacity for information to be spread on social media likely exceeds the capacity of any other medium in history (e.g., Lu, Wen, & Cao, 2014).

The rapid spread of content has led to a social-media environment in which moralized content is ubiquitous. Approximately 90% of a social-media users report seeing at least "a little" political content in their social-media feeds (Duggan & Smith, 2016), and online platforms are now one of the primary sources of morally relevant stimuli people experience in their daily life (Crockett, 2017). The spread of moralized content can have important consequences in the domain of morality and politics, such as in the case of "online firestorms," or massively cascading bursts of moral outrage that ruin the reputation of individuals or organizations within hours (Pfeffer, Zorbach, & Carley, 2014; Ronson, 2016; Rost, Stahel, & Frey, 2016), and viral prosocial campaigns that raise millions of dollars in

only a few days (Van Der Linden, 2017). Thus, it is important to understand the psychological processes that underlie the spread of moralized content online.

In this article, we propose a model for understanding the psychological underpinnings of social contagion in the domain of moral and political discourse online. First, we review recent evidence documenting that the expression of moral emotions plays an important role in the spread of moralized content online (a phenomenon we call *moral contagion*). Second, we propose a model of moral contagion called the motivation, attention, and design (MAD) model that explains why such content spreads, drawing on insights from a diverse range of psychological theory and data. The MAD model proposes that people are motivated to share moral-emotional content based on their group identity, that such content is especially likely to capture attention, and that the design of social-media platforms interacts with these psychological tendencies to further facilitate its spread. We review evidence in support of each model component (as well as their interactions) and propose several testable hypotheses that can spark progress in the scientific investigation of social-contagion processes that may underlie civic engagement and activism, political polarization, propaganda and disinformation, and other moralized behaviors in the digital age.

Moral Contagion in Online Social Networks

In this section we review recent evidence that sheds light on what factors affect the spread of content across various online contexts. In other words, we ask what makes different types of online content go "viral." Two key findings emerge: The first is that emotionally arousing content is associated with increased sharing across various online contexts; the second is that, in the specific context of moral and political discourse, moral-emotion expression may play an important role in the spread of content (a phenomenon we call moral contagion). Taken together, these factors help to identify the type of content that is more likely to spread within social networks, especially in online settings.

Emotionally arousing content is likely to be shared

People tend to share emotional experiences with others (Rime, Mesquita, Philippot, & Boca, 1991). For instance, people are more likely to share social memories, tell stories about themselves, and pass on urban legends if they are emotionally arousing (Christophe & Rimé, 1997; Heath, Bell, & Sternberg, 2001; Peters & Kashima, 2007). Sharing emotional experiences may be a functional tool for increasing social bonding: When people share emotional experiences with others, it leads to perceptions of similarity, emotional convergence, and greater coordination during goal-directed action (Locke & Nekich, 2000; Peters & Kashima, 2007). It can also serve the function of signaling important elements of one's social identity or social norms to their social community, which may increase their stature within the community (Jordan & Rand, 2020). Thus, the expression of emotion seems to serve a number of important functions in communities.

Recent work investigating the spread of online content across various domains suggests that emotionally arousing content is robustly associated with increased sharing. One study of 6,956 popular news articles found that articles that induced high-arousal emotions, including awe, anger, and anxiety, were more likely to be shared via e-mail (Berger & Milkman, 2012). A larger study of 65,000 news articles across various languages replicated these basic findings, although the specific emotions associated with sharing varied across cultures (Guerini & Staiano, 2015). In the case of social media, multiple studies have documented that emotional content is associated with increased sharing on various platforms, including Facebook (Heimbach, Schiller, Strufe, & Hinz, 2015; Kramer, Guillory, & Hancock, 2014), Twitter (Hansen, Arvidsson, Nielsen, Colleoni, & Etter, 2011; Quercia, Ellis, Capra, & Crowcroft, 2011; Stieglitz & Dang-Xuan, 2013), Google+ (Heimbach et al., 2015; Hochreiter & Waldhauser, 2014), and Weibo (a popular Chinese microblogging platform; Fan, Zhao, Chen, & Xu, 2014). Thus, the human tendency to share emotional experiences carries over to numerous online platforms.

Moral-emotional content is likely to be shared

In the context of political discourse on social media, the combination of moral and emotional expression may be particularly important for sharing. For instance, political discussions infused with emotional language were shared the most widely in a study investigating discourse related to an election (Stieglitz & Dang-Xuan, 2012). Political news articles framed in terms of morality and that included emotional language were the most widely shared across Facebook and Twitter (Valenzuela, Piña, & Ramírez, 2017). Furthermore, a study investigating moral and political discourse on Twitter using more than 500,000 messages discussing multiple contentious political topics found that expressions of moral emotion were most associated with sharing—even more

Table 1. Sample Tweets From Each Political Topic Separated by Ideology

Topic and mean ideology of retweeters	Twitter message
Gun control	
Conservative	America needs to Arm itself. Stand and **Fight** for Your Second Amendment Rights. We are literally in a **War** Zone. Carry and get Trained.
Liberal	Thanks to **greed**, the republication leadership & the #NRA – No one is **safe** #SanBernadino #gunsense #guns #morningjoe
Same-sex marriage	
Conservative	Gay marriage is a diabolical, **evil** lie aimed at **destroying** our nation. #o4a #news #marriage
Liberal	New Mormon Policy Bans Children Of Same-Sex Parents-this church wants to **punish** children? Are you kidding me?!? **Shame** https://. . .
Climate change	
Conservative	Leftists take 'global warming' based on **bad** science as **faith** and act on it, but proven voter fraud is just racism #tcot #teaparty
Liberal	**Fighting** #climatechange is **fighting** hunger. Put your #eyesonParis for a fair climate deal.

Note: Moral-emotional words are in bold. Adapted from Brady, Wills, Jost, Tucker, and Van Bavel (2017), in which moral-emotional language was measured as words that co-occurred in two existing lexicons: the Moral Foundations Dictionary developed to measure language that references the domain of morality (Graham, Haidt, & Nosek, 2009) and the Linguistic Inquiry and Word Count lexicon developed to measure language that is emotional (Tausczik & Pennebaker, 2010). This choice was theoretical and was based on the idea that words co-occurring in both lexicons should represent emotional language that is associated with how people discuss morality. Brady et al. (2017) also empirically validate the formation of these categories by having participants judge words and tweets. One limitation of this method is that it does not capture all potential moral-emotional words because measurement is limited to words that appear in the lexicons. For instance, the conservative same-sex marriage tweet contains the word "diabological," which could be considered moral-emotional on the basis of the theoretical definition provided above. However, this word is missed by the lexicon method used in Brady et al. (2017). This limitation can be addressed by using newer methods in natural-language processing and machine learning to measure specific moral emotions by combining theory-driven feature selection and data-driven approaches that capture how language is used on Twitter. There are also available methods for expanding existing lexicons with data-driven approaches (see Frimer, Boghrati, Haidt, Graham, & Dehgani, 2019).

consistently expressions that included only either moral or emotional language (Brady, Wills, Jost, Tucker, & Van Bavel, 2017). In fact, every moral-emotional word added to a tweet was associated with a roughly 20% increase in sharing on average (Brady et al., 2017; for examples of moral-emotional language, see Table 1). The association between moral-emotional language and sharing was replicated during the 2016 U.S. election campaigns among political leaders, where retweets of messages from more than 500 presidential candidates and members of the U.S. Congress were analyzed (Brady, Wills, Burkart, Jost, & Van Bavel, 2019). In the context of political discourse online, moralized content containing moral-emotion expression is consistently associated with increased sharing across various topics, among laypeople and political leaders, and during consequential political events. We call this phenomenon *moral contagion*.

Moral contagion defined

Moral contagion refers to the idea that moral-emotion expression is associated with the spread of moralized content in online networks. The strongest form of moral contagion would suggest that moral-emotion expression is both necessary and sufficient for the spread of moralized content online. However, this is unlikely to be true because deciding to share content online, like all human behavior, is a multifaceted process. Furthermore, there are likely to be context-sensitive communication norms that moderate which specific emotion expressions affect diffusion (Brady et al., 2017; Postmes, Spears, & Lea, 2000). In 2015 when the U.S. Supreme Court legalized same-sex marriage, positively valenced emotion expressions were associated with the greatest diffusion in the context of supporting the court ruling rather than sanctioning expressions such as outrage (Brady et al., 2017). Further, outrage expressions that sanction the political out-group might spread widely in networks in which out-group derogation is normative, but in other networks or communication contexts this may not be the case (see, e.g., Reicher, 1984). More accurately, then, moral contagion implies that moral-emotion expression facilitates the spread of moralized content online, and in contexts of moral and political communications it is on the average a highly significant factor.

The concept of moral contagion consists of two key components: the spread of content containing moral-emotion expressions and a specific social-contagion process based on information diffusion. Moral emotions are associated with appraisals, eliciting conditions, and functions that are specifically tied to the context of

morality (e.g., moral outrage, contempt, moral disgust, shame, elevation; Haidt, 2003; Haidt, Rozin, McCauley, & Imada, 1997; Hutcherson & Gross, 2011). Here we focus on the differences between moral- and nonmoral-emotion expressions. In the context of social-media communications, the emotion expression represented in a message (which is known to any user who sees the message) is ostensibly more important than the underlying emotional state of the message author (which is not known to any user who sees the message) in terms of affecting other people in an online network.

Psychological-constructivist (Barrett, 2013) and social-functionalist (Keltner & Haidt, 1999) accounts of emotion expression help distinguish between moral- and nonmoral-emotion expressions. Psychological-constructivist accounts of emotion define emotion expression in language as the usage of specific culturally and contextually defined concepts that represent underlying feelings defined by valence and arousal, or "affect" (Lindquist, MacCormack, & Shablack, 2015). Social-functionalist accounts distinguish emotions on the basis of the different social functions they serve, such as signaling specific social information to others. Building from these two accounts, we define moral-emotion expression in social-media text as representational expressions of affect that reliably signal, either to others or to the self, that something is relevant to the interests or good of society, as defined by the conceptual knowledge of the expresser. For instance, moral-outrage expression is a prototypical moral-emotion expression because it normally indicates that the expresser perceives some transgression against one's concept of right and wrong has occurred (e.g., Rozin, Lowery, Imada, & Haidt, 1999; Tetlock, Kristel, Elson, Green, & Lerner, 2000). On the other hand, the expression of sadness is not a prototypical moral-emotion expression because expressions of sadness have a much wider range of eliciting conditions and contextual cues that may have nothing to do with morality (e.g., the death of a pet because of old age). Thus, the inference that something morally relevant has occurred when one expresses sadness is less likely to be accurate. However, according to our definition, a cultural or message context in which the shared concept of sadness more reliably represented moral relevance could change the status of sadness as an instance of moral-emotion expression for that specific context.

Moral-emotion expressions are also those that reliably signal to the self that something morally relevant has occurred, such as in the case of shame or guilt when expressions can be used to guide one's future behavior (Clore & Huntsinger, 2007). Indeed, morality is central to our understanding of identity (Aquino & Reed, 2002; Strohminger & Nichols, 2015). Inherent in the concept of moral contagion, then, is the idea that moral-emotion expressions are among the most powerful signals to the self and others about one's identity. As such, they may be among the most functionally relevant forms of expressions in the context of moral and political discourse online, where moral and political identities are salient.

The other key component of moral contagion is *contagion*, which refers to the spread of moral-emotional content through online sharing that takes the form of information diffusion (Bakshy et al., 2012). The object being spread is the symbolic representation (language, images) of the moral-emotion expression (information). Exposure to the information can serve as input to one's own evaluation of the object or event in question. For an illustration of this process (see Fig. 2), consider an example in which Nancy, a partisan, evaluates political out-group Senator John Doe after logging onto Twitter. Nancy views a message from a user in her social network stating that "John Doe is the worst," and their message also includes a vomiting emoji. From Nancy's perspective, the message sent by the user is a representation of their negative emotions felt toward John Doe and provides Nancy with information about how other people in her network evaluate John Doe. This information (how other people feel about John Doe) is input into Nancy's own evaluation of John Doe and can guide her subsequent behavior and/or her emotions. One quick and direct route to action in this context is to simply retweet the message, thereby displaying Nancy's updated evaluation that was influenced by the message and spreading the information further across the social network.

It is possible that Nancy retweets a message and at the same time experiences an emotional state while typing such as the emotion represented in the original message. Indeed, this is likely often the case. For Nancy to retweet a message, however, it is not necessary for the emotional state represented in the original message to be experienced fully or at all. If it affects her behavior such that she retweets the message, then contagion has occurred because the information in the message has now been further spread in the network. The process of moral contagion is more concerned with the spread of the emotional information through social networks, which occurs via sharing and posting behaviors. Note that in the context of social media, the only impact a user can have on others in their social network is through the messages that represent emotion (and not the offline emotional state of the user, which may or may not be aligned with the emotional expression).

Our use of the term contagion here as information diffusion departs from some traditional uses of the term *emotional contagion* that refer to the spread of emotional

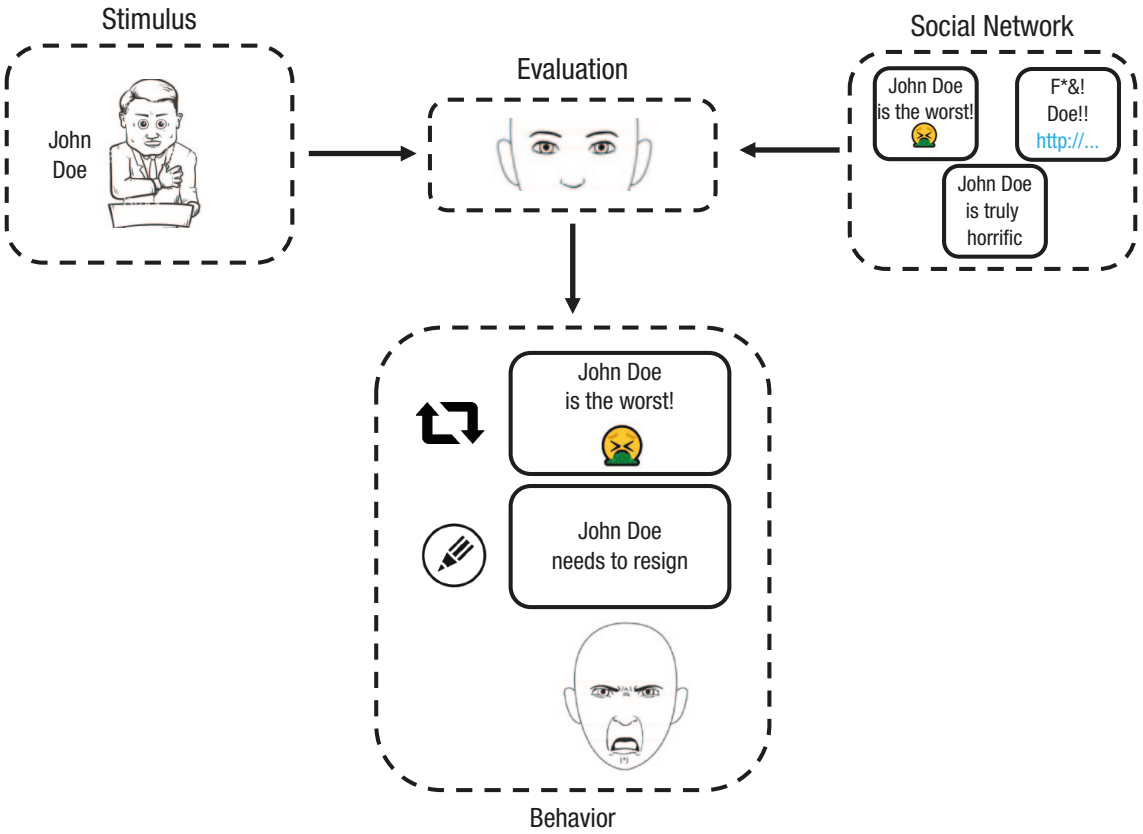

Fig. 2. Illustration of the process of moral contagion via social appraisal in online networks. Both the stimulus or event in question and other people's reactions in one's social network to that stimulus (which may constitute social norms) serve as information input in one's evaluation of the stimulus. One's evaluation may be influenced by others' reactions and may lead to online behaviors and/or feelings that are similar to others' reactions.

states because of automatic muscle mimicry of facial expressions during in-person interactions (Hatfield, Cacioppo, & Rapson, 1993). Of course, emotional contagion in this traditional sense is less relevant on social media, where physical face-to-face interactions do not occur, and thus contagion based on an evaluation process (Fig. 2) is necessitated by the context. Furthermore, whereas emotional contagion during face-to-face interactions is thought to be automatic, in the context of social media people may be carefully constructing these messages, and the decision to share them may be deliberate for many people.

The definition of contagion used here is compatible with social-appraisal theory, which supposes that emotions spread as the result of people factoring in others' emotional reactions into their appraisal of a stimulus or event (Fischer & Manstead, 2004). An appraisal is a rapid evaluation of a stimulus or event that serves as information to the organism about whether the stimulus/event is relevant to its well-being (Arnold, 1960; Lazarus, 1966), and specific appraisal combinations may lead to specific emotions (Ellsworth & Scherer, 2003). Stated in terms of social-appraisal theory, then, moral contagion is the spread of moralized content as a result of people incorporating others' moral-emotional expressions as informational input into their own appraisal of a situation, which can guide their decisions in sharing the content on social media and inform their own emotional state. Such moralized appraisals can have a range of effects on our evaluations—for example, thinking in extremes and social conflict (Luttrell, Petty, Briñol, & Wagner, 2016; Skitka, Bauman, & Sargis, 2005; van Bavel et al., 2012).

Summary

In this section we reviewed evidence regarding the consumption and sharing of news headlines, moral and political discourse, and crafted messages sent by political leaders suggesting that moral-emotion expression plays a key role in the diffusion of moralized context online (a phenomenon we call moral contagion). We defined moral-emotion expression as emotion expression that reliably signals that an object or event is relevant to the interests or good of society from the perspective of the expresser. We defined moral contagion as the diffusion

of moralized information on the basis of social-appraisal processes.

The phenomenon of moral contagion has a number of important implications for understanding morality and politics in the digital age. First, moral contagion may be an antecedent process feeding into polarization or the increased distance among members of different political ideologies in online social networks (Barberá, Jost, Nagler, Tucker, & Bonneau, 2015; Brady et al., 2017) that people across the political spectrum have argued is a significant threat to civil discourse (Cushman, 2017; Mrowicki, 2015; Noel, 2017). Second, the process of moral contagion highlights that social-media posts containing moral-emotion expression receive increased amounts of positive social feedback on social media, which has the potential to amplify our natural tendency to express moral emotions such as outrage via social reinforcement (Crockett, 2017; see below). Third, moral contagion describes a concrete process by which social leaders, including activist organizations or political elites, can gain massive exposure for their ideas and signal social norms (Brady, Wills, et al., 2019), which has important consequences for social influence in the digital age (Pärnamets, Reinero, Pereira, & Van Bavel, 2019). Fourth, the use of moralized language online may even precipitate political action, including violence in the real world (Mooijman, Hoover, Lin, Ji, & Dehghani, 2018). To better understand these important phenomena, we propose a new model called the MAD model that clarifies some of the key psychological processes that underlie moral contagion. This model can guide future research to help understand the spread of moralized content online and its associated consequences (see Fig. 3).

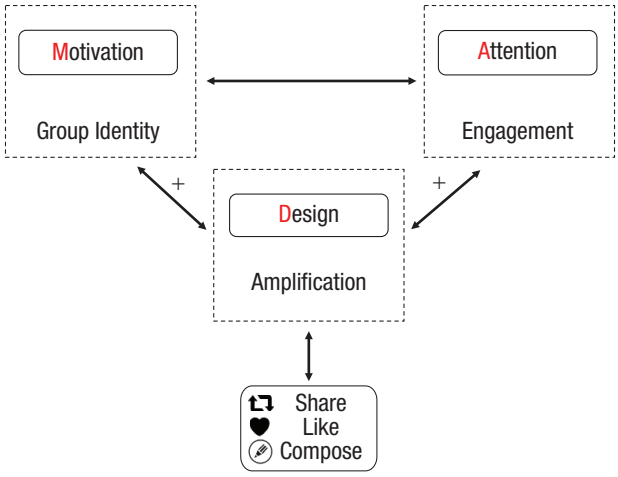

Fig. 3. The motivation, attention, and design (MAD) model of moral contagion. The MAD model helps to explain why moralized content spreads online by considering the interaction of our psychology and the social-media environment in which we interact: group-based motivations, how moralized stimuli engage our attention, and how the design of social media amplifies these elements of our psychology. Each component interacts to produce the ultimate decision to share or post moralized content in social networks.

Group-Identity Motivations and the Spread of Moral-Emotional Content

In this section we examine the social motives—the goals, desires, and wants—that are likely to play a role in the spread of moral-emotional content online. Social media, as its name suggests, is fundamentally characterized by repeated social interactions. As in other social contexts, individuals' behavior is largely motivated by a desire to feel a sense of belonging in their social networks (Baumeister & Leary, 1995). As we argue below, individuals' behavior on social media is dominated by additional social motives, especially the desire to maintain or enhance their social status in relation to a specific group identity.

Whether it is political discussions that resemble echo chambers and often highlight political group differences (Barberá et al., 2015; Brady et al., 2017), political leaders who disseminate partisan content (Brady, Wills, et al., 2019), or the viral spread of fake political news (Lazer et al., 2018; but see Guess, Nagler, & Tucker, 2019), for some users social media serves as a constant reminder of our political group identities. In fact, almost all social-media users report that when they log onto social media they see at least "a little" political content (94% on Facebook and 89% on Twitter; Duggan & Smith, 2016). Furthermore, a recent analysis suggests that people are significantly more likely to encounter morally or politically relevant information that makes them outraged on social media compared with other sources (Crockett, 2017). During the exchange of moralized political information like that which is found in exaggerated amounts on social media, our political group identities (e.g., liberal vs. conservative) are likely to be hypersalient (B. Simon & Klandermans, 2001). Put another way, for many users simply logging on to social media can serve as a group-identity manipulation.

The high salience of group identity in the social-media environment has important implications for the form and function of information exchange as it pertains to moralized content specifically. Below, we leverage social-identity theory (Tajfel & Turner, 1979), self-categorization theory (Turner et al., 1987, 1994), and intergroup emotions theory (E. R. Smith, Seger, & Mackie, 2007) to outline specific group-based motivational contexts that we argue should be associated with corresponding sets of group-based emotion expressions useful for fulfilling identity-based motivations. We argue that this broad "social-identity approach" offers a

Table 2. Summary of Types of Group-Identity Threats, Expected Emotional Responses, Proximate Functions, and Corresponding Example Messages From Social Media

Group-level elicitor	Moral-emotional response	Proximate function	Example social-media message
Threat from out-group	Outrage, contempt	Out-group derogation	Liberals' attempts to make Trump look bad are just despicable! #MAGA
Threat from out-group	Gratitude, elevation	Group affirmation	We are grateful for Blasey Ford's courage. #IBelieveHer #BlueWave2020
Threat from in-group	Guilt, shame	Reparation, denial	I acknowledge my role in misogyny culture and will call out other males. #HowWillIChange

powerful framework for understanding the expression and diffusion of moral emotions online.

Intergroup-identity-based motivations

According to social-identity theory and self-categorization theory (Tajfel & Turner, 1979; Turner, Hogg, Oakes, Reicher, & Wetherell, 1987), when group memberships are highly salient (as they are on social media), people's individual identities become subsumed by group identity. In this case, people's attitudes, emotions, and behaviors are influenced by evaluations made in terms of group, rather than individual, goals (Abrams & Hogg, 2004; Tajfel & Turner, 1979). Through the lens of these group evaluations, people perceive themselves in terms of group characteristics and view in-group members as more similar and out-group members as more different (Hornsey, 2008). This shift from the self to group identity is associated with a motivation to maintain positive distinctiveness between the in-group and relevant out-groups (Brewer, 1979; Tajfel & Turner, 1979; Turner, Oakes, Haslam, & McGarty, 2007), belong to the group (Jetten, Branscombe, Schmitt, & Spears, 2001), and reduce uncertainty (Hogg, 2007). On social media, then, engagement with in-group members can reinforce group boundaries and fulfill a number of important group-identity motives, including positive distinctiveness, the need to belong, and epistemic needs.

Our central argument is that expressions of emotion during moral and political communications play an important role in fulfilling these group-identity motives. Emotions are functional responses that regulate behavior to help individuals achieve their goals (Frijda, 1986; Keltner & Haidt, 1999), and emotions can also regulate behavior on the basis of group-level goals. The central insight of intergroup emotions theory (Mackie, Silver, & Smith, 2004) is that when group identities are salient people begin to experience and express emotions on the basis of elicitors that are relevant to the group. For instance, if someone identifies as a Democrat and Democrats are attacked by conservative media, that person may experience a negative emotional response as if they were attacked personally. Like individual-level emotions, group-based emotions are functional: They regulate behavior in ways that help people achieve group-based goals, including the fulfillment of social-identity motives (E. R. Smith et al., 2007).

We propose on the basis of social-identity theory and intergroup emotions theory that posting or sharing moralized content containing moral-emotional expressions on social media helps people satisfy the motivation to maintain a positive group image, which broadly satisfies a number of identity motives (see also Van Bavel & Pereira, 2018). Because there are various routes through which in-group members can maintain a positive group image in different contexts (Hornsey, 2008; Tajfel & Turner, 1986), our model considers two key contexts discussed in the literature that can frustrate goals to maintaining a positive group image: identity threat originating from out-group members and identity threat originating from in-group members (See Table 2). We propose specific moral-emotional expressions that help to reestablish a positive group image in the face of group-identity threats.

When out-group members pose threats to the moral values of the in-group, out-group derogation is a common in-group response to uphold a positive in-group image—although it depends on the norms of the group and the strength of group identification (Branscombe, Ellemers, Spears, & Doosje, 1999). In other words, condemning an out-group's behavior makes one's in-group appear better by comparison. When a threat from an out-group is present, the expression of moral emotions that sanction people or behaviors by signaling disapproval function to derogate the out-group through their expression. For example, outrage is an emotion that consists of feelings of anger and disgust, as well as specific cognitive and behavioral tendencies associated with blame and punishment (Salerno & Peter-Hagene, 2013; Tetlock et al., 2000). When group moral values

are threatened, people automatically respond with moral outrage directed toward the source of the threat (Tetlock et al., 2000). Relatedly, group-level anger increases verbal and physical confrontation with out-groups (Iyer, Schmader, & Lickel, 2007; Mackie, Devos, & Smith, 2000; E. R. Smith et al., 2007). Contempt—an emotion that consists of negative feelings based on feelings of moral superiority (Rozin et al., 1999)—is another emotion that can serve out-group derogation. For example, contempt is associated with maintaining social and political group hierarchies (Miller, 1997; Rozin et al., 1999) and feelings of moral superiority over other groups (Ekman, 1994; Izard, 1977). Thus, expressions of outrage and contempt may help to maintain a positive group image in response to group threat by derogating the out-group.

On social media, threats to group values originating from the out-group often take the form of informational content documenting value-violating actions from political out-group members (e.g., news articles) and, to a lesser extent, direct interaction with rival political group members. For instance, more than 100 million Americans were exposed to political ads paid for by Russian agents that explicitly sanctioned and highlighted bad behavior of political candidates and political groups (Allcott & Gentzkow, 2017; Timberg, Dwoskin, Entous, & Demirjian, 2017). The available responses to threatening content online is the posting or sharing of commentary on the content by using, for example, outrage expression that derogates out-group members. On the other hand, direct confrontation of out-group members often occurs during arguments in the comment sections of a user's post (Facebook) or through direct replies (Twitter) with anger and outrage expression directed at the out-group member. For instance, in response to polarizing events, political out-group members on Twitter sometimes talk directly to one another via the reply feature, and it largely consists of a buildup of anger expression and a subsequent rise in in-group identification (Yardi & Boyd, 2010).

Another response to threats to group values originating from out-groups is the affirmation of group values as well as strong displays of group affiliation (Ellemers, Spears, & Doosje, 2002). The expression of moral emotions that promote positive in-group evaluations readily serve group affirmation. The moral-emotion gratitude—a positive emotion in response to the perception of a good deed that was directed toward oneself (McCullough, Emmons, Kilpatrick, & Larson, 2001)—can affirm group identities. Gratitude is associated with positive evaluations of others, intentions to form bonds with them, and even costly behaviors that benefit the original benefactor (Algoe, Haidt, & Gable, 2008; DeSteno, Bartlett, Baumann, Williams, & Dickens, 2010). On social media, we have seen the affirmation of political identities associated with gratitude expression, as in the case of gratitude expressed toward Christine Blasey Ford by Democrats during the polarizing Brett Kavanaugh confirmation hearings in September 2018. Thousands of women took to Twitter to express gratitude to Blasey Ford—who testified that U.S. Supreme Court nominee Brett Kavanaugh committed sexual assault—of her bravery and also condemned Republicans for supporting Kavanaugh (Penrose, 2018).

The moral-emotion elevation—a pleasant feeling after witnessing virtuous behavior (Haidt, 2000)—also serves group affirmation. For instance, elevation is associated with positive evaluations of others and increased social affiliation (Haidt, 2000). During the 2016 election, expressions of positive evaluations toward American troops using language such as "hero" and toward religious organizations using language such as "faith" were shared widely when posted by conservative political leaders (Brady, Wills, et al., 2019). The expression of moral emotions such as gratitude and elevation are examples of emotions that serve group affirmation in response to in-group threat: They motivate positive evaluations of one's group and increase in-group affiliation behaviors. These expressions can also signal identity to others online, which can increase support and promote affiliation.

Contexts in which one's group identity is under threat as a result of the behaviors of their own group members (in particular behaviors toward a lower-status out-group) often leads group members to engage in behaviors to repair the in-group's image (Doosje, Branscombe, Spears, & Manstead, 1998). The expression of "self-conscious" moral emotions (Tracy & Robins, 2004) that motivate people to engage in interpersonal (and intergroup) reparative actions facilitates the process of group-image reparation. Guilt—a negative emotion experienced when focusing on the negative behaviors of one's self or group (Tracy & Robins, 2006)—is one emotion that serves reparation behavior. Guilt is associated with internal attributions that lead to behaviors, including apologies, confessions, and prosocial actions (de Hooge, Zeelenberg, & Breugelmans, 2007; Niedenthal, Tangney, & Gavanski, 1994; Sheikh & Janoff-Bulman, 2010; Tangney, Miller, Flicker, & Barlow, 1996; Tangney & Tracey, 2012). Collective guilt drives the willingness to compensate for bad in-group behavior and attitudes toward reparations, even if one did not perform the behavior themselves (Brown, González, Zagefka, Manzi, & Ćehajić, 2008; Doosje et al., 1998; Wohl, Branscombe, & Klar, 2006). By motivating conciliatory behaviors, guilt can promote actions required to repair a group's image, even if one was not the group member who misbehaved. This can manifest in online messages that apologize for

fellow in-group members or call their behavior into question.

On the other hand, a positive group image could also be maintained in the face of threat originating from in-group behavior by devaluing the dimension that is threatening (i.e., attempting to downplay the extent to which in-group behavior violated values; Hornsey, 2008). The self-conscious moral-emotion shame—a negative emotion experienced when focusing on the negative events in terms of one's self or group image (Tracy & Robins, 2006)—is one emotion that serves such protective, devaluation behavior. Shame is associated with external attributions and is linked to blaming others, distancing oneself from the negative event, and denial (Brown et al., 2008; Niedenthal et al., 1994; Tracy & Robins, 2006). Group-level shame elicited by in-group behavior is driven by a desire to maintain a positive reputation for the in-group (Brown et al., 2008). This is different from attempts to "shame" out-group members, which is very common online and often involves attempts to present public evidence of wrongdoing from out-group members or spark collective action to condemn (Jacquet, 2016).

Users on social media often bring to light behaviors of one's group that are counternormative, such as sexual-assault allegations against liberal men that were made public on social media as part of the #MeToo movement. Some liberal men responded with guilt expression such as open statements of apology and promises to become more aware of difficulties faced by women. The hashtag #HowWillIChange went viral rapidly and was adopted by men across the world in response to #MeToo (Vaglanos, 2017). On the other hand, a viral meme in response to various instances of misogyny has been the phrase "not all men" from those who deny that the typical man should be blamed for misogyny in American culture (Fordy, 2014).

Intragroup-identity-based motivations

Thus far we have argued that group-based moral-emotional expressions satisfy the identity-based motivation to uphold a positive in-group image relative to out-groups. This explanation describes a motivation that essentially pertains to intergroup relations. At the same time, group-based moral-emotional expressions can also satisfy an identity-based motivation that fundamentally pertains to within-group relations: the need to maintain an image as a good group member in the eyes of other group members. In other words, expressing moral emotions that derogate the out-group or bolster the in-group can enhance one's reputation and increase group belonging. Insofar as expressing moral emotions during moral and political communications is favored by observers in one's social network (an idea supported by the increased positive social feedback associated with posting such content; Brady et al., 2017), it follows that such behavior can increase one's social reputation.

Research on the reputational benefits of specific types of moral decisions supports the argument that expressing moral emotions in the context of moral and political communications can enhance one's reputation within their group. For instance, when people make deontological moral decisions that are often associated with emotion-based processes (e.g., Greene, Nystrom, Engell, Darley, & Cohen, 2004), they are viewed by others as more moral and more trustworthy (Everett, Faber, Savulescu, & Crockett, 2018; Everett, Pizarro, & Crockett, 2016; Rom & Conway, 2018; Uhlmann, Zhu, & Tannenbaum, 2013). Furthermore, when people express outrage by punishing others who propose unfair offers in economic games, they are viewed as more trustworthy by others (Jordan, Hoffman, Bloom, & Rand, 2016; Jordan & Rand, 2017). These data suggest that in moral contexts, showcasing one's emotional reactions can increase their status—as long as those reactions are well aligned with the value system of their social network.

Expressing one's moral and political attitudes with moral emotions may be akin to punishing unfair agents in offline social interactions. By expressing attitudes with moral emotions, one is signaling clearly that they endorse, if not share, the relevant attitudes with their social group. For example, in online settings people are more likely to express outrage toward policies they oppose when their identity is *not* anonymous, suggesting that the opportunity to signal to others should be associated with a greater likelihood of expressing outrage online (Rost et al., 2016). Given that moral emotions may be expressed at a much higher rate on social media compared with other media (e.g., Crockett, 2017), the expression of moral emotions can also signal a sense of shared understanding of communication norms. Online environments rapidly create new social norms for communication that vary by networks and influence individual communication styles (Postmes et al., 2000). The more frequently moral-emotion expression is used online in a network, the more it can be used to signal shared understanding of communication norms in that network, thereby demonstrating social value. Relatedly, people feel a shared identity with others who are expressing similar emotions (Livingstone, Shepherd, Spears, & Manstead, 2016), and thus as the norm to express emotions increases, group identification may become even stronger.

In summary, recent evidence supports the idea that by expressing moral emotions targeted at out-group members or supporting in-group members can enhance

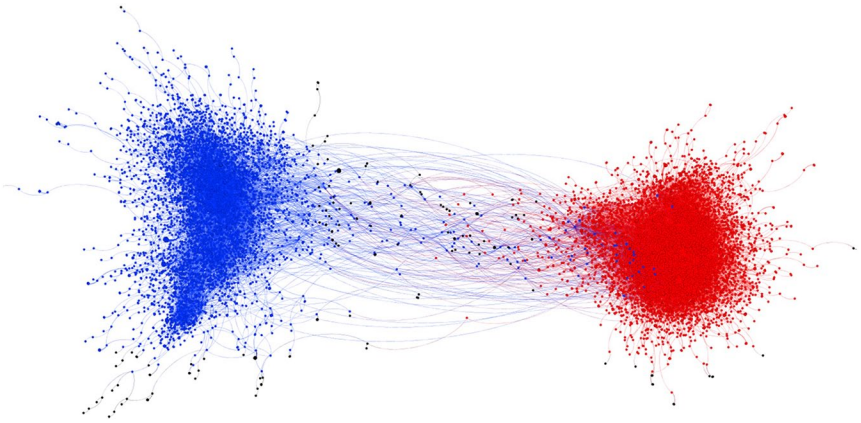

Fig. 4. Network graph of moral contagion shaded by political ideology. The graph depicts messages containing moral and emotional language, and their retweet activity, across all political topics (gun control, same-sex marriage, climate change). Nodes represent a user who sent a message, and edges (lines) represent a user retweeting another user. The two large communities were shaded on the basis of the mean ideology of each respective community (blue represents a liberal mean; red represents a conservative mean). Adapted from Brady et al. (2017).

one's reputation, which satisfies the social motivation to maintain good standing in one's group. This could help to explain the rise of largely symbolic outrage expressions in response to moral transgressions in online social networks that appear to be driven by a desire for social recognition (see, e.g., Johnen, Jungblut, & Ziegele, 2018), as well as recent evidence suggesting that politically active people are likely to express morality for social-status motivations (Grubbs, Warmke, Tosi, & James, 2019).

Predictions and implications

One basic tenet of the social-identity approach is that people who identify more strongly with their group in a given context are more likely to behave in ways driven by social-identity processes (Tajfel & Turner, 1979). This raises two key predictions regarding the spread of moralized content on social media. First, people who identify more strongly with political groups (e.g., Democrats and Republicans) should be the most likely to post and share moral and political content that uses moral-emotional expression. Second, over and above existing characteristics of a person (e.g., their average level of group identification), the impact of social identity should be heightened during contexts of intergroup conflict and threat—such as periods with intense group polarization. Such contexts will push people to behave in ways based on group-identity motivations, which in the context of social media can lead to the greater use or sharing of moral-emotional messages. In other words, the MAD model predicts that during contexts of high group-identity threat, the frequency of posts and shares containing moral-emotional expressions should be greater than contexts with lower threat. For instance, if a political group faces a serious threat of losing power in an election, or recently lost power in an election, it is likely that partisans will respond by increasing the posting and sharing of moral-emotional content. The specific types of moral-emotional expressions that increase in social-media messages depends on the specific manner in which the threat is construed and perceived (see Table 2). Note that this pattern of behavior could result in communications that are largely bound by in-group bias, or in other words the sharing of information by one's in-group rather than out-group (Brady et al., 2017, see Fig. 4).

One important insight from social-identity theory is that group identification has its effects on behavior by increasing conformity to specific group norms (Reicher, 1982, 1984). Thus, an important moderating variable for the predictions proposed above is the nature of specific group-communication norms present in a social network. Specifically in computer-mediated environments, people rapidly change their communication strategy on the basis of evolving dynamic communication norms (Postmes, Spears, & Lea, 1998; Postmes et al., 2000); see also discussion below). From these insights, the MAD model predicts that the posting and sharing of content containing moral-emotional expressions should be the most heightened in contexts in which group identification is high (e.g., when a political group is under threat) and in which the sanctioning of

out-groups via outrage expression is highly normative (e.g., in social networks in which ideological extremity is high).

Finally, we predict that leaders can leverage features of social identity to drive the feelings and actions of followers—a feature of identity leadership (Haslam, Reicher, & Platow, 2010). The use of moral-emotional expression among political leaders on social media is important because their messages reach large audiences and often drive news coverage. Recent evidence supports the idea that political leaders leverage moral-emotional expressions and that these expressions are associated with increased online engagement from their constituents (Brady, Wills, et al., 2019).

The predictions raised here by the MAD model, which focus on how emotion expressions in the context of information diffusion function to maintain group-identity goals, dovetail nicely with recent process models of group-based emotion regulation (Goldenberg, Halperin, van Zomeren, & Gross, 2016). These models propose that people have group-identity goals and use them to regulate the emotions of other group members. Indeed, when people express moral emotions such as outrage to derogate the out-group and protect their group identity from threat, they may be instrumentally upregulating their emotional state (see Goldenberg et al., 2016, p. 128) or expression to motivate other group members to participate in the outrage expression that will further diminish the out-group's image (see also Zaki & Williams, 2013). Combining the idea of group-emotion regulation with the MAD model leads to the specific prediction that when one's group identity is under threat people will deploy emotion-regulation strategies that increase their own and others' moral-emotion expressions in social-media messages. One interesting possibility is that this process might affect communication norms in an online social network over time, thus making the use of explicit upregulation strategies less necessary when the expressions of emotions such as outrage are more widespread.

The motivation component of the MAD model presented here has important implications for understanding the psychology of political debate on social media, which some have likened to a "dumpster fire" (Maza, 2018) and often represent partisan bias more than thoughtful and balanced discussions of issues (Bakshy, Messing, & Adamic, 2015; Barberá et al., 2015; Brady et al., 2017; Colleoni, Rozza, & Arvidsson, 2014; Himelboim, Mccreery, & Smith, 2013). Although some research would suggest that a solution to better civil discourse online would be to increase people's exposure to ideologically diverse viewpoints (Grönlund, Herne, & Setälä, 2015; Mutz, 2002), existing data demonstrate that exposure to ideologically diverse viewpoints on social media can actually backfire—*increasing* political polarization (Bail et al., 2018; Yardi & Boyd, 2010). These findings are readily explained by the MAD model: Exposure to out-group political views may simply make one's in-group identity more salient and increase the possibility that people respond with moral-emotion expressions that differentiate the groups by derogating the out-group or highlighting in-group virtues. Such outcomes will be exacerbated if the individuals are embedded within online social networks in which communication norms are more likely to condone lashing out and extreme language rather than thoughtful reflection.

We also discussed how moral-emotion expressions serve intragroup-identity motivations, and this function has important implications for understanding the psychology of political polarization as it unfolds in online contexts. If people express moral emotions that sanction out-groups (e.g., outrage) for reputational benefits, either consciously or unconsciously, then in-group and out-group differences may appear worse online than they are in reality (e.g., "false polarization"; Levendusky & Malhotra, 2016). For example, a Democrat viewing online discussions among Republicans who have norms of outrage expression may perceive extreme levels of disagreement with Democratic policy or politicians, when in reality each individual Republican does not disagree with Democrats to the same extent as it appears. Thus, one important question for future research is whether the spread of moralized content infused with moral emotions can dynamically increase conformity to people's perception of group norms online, even when group members' attitudes are not as extreme as their public expression would entail (similar to cases of pluralistic ignorance; Prentice & Miller, 1993). This potential consequence of moral contagion may be especially important because norms strongly influence many aspects of behavior (Cialdini, Kallgren, & Reno, 1991).

Summary

In this section, we argued that social media is a context in which our political group identities are hypersalient. As a result, people are strongly motivated to maintain not only a positive in-group image relative to the out-group (intergroup-identity motivation) but also a positive reputation of themselves in their group (intragroup-identity motivation). Expressions of group-based moral emotions readily serve these motivations by derogating the out-group (e.g., outrage, contempt), bolstering the in-group (e.g., elevation, awe), and repairing the in-group's image (e.g., guilt, shame). These group-identity processes help explain why moral-emotional expressions are often shared widely during moral and political discussions on social media, where our political group

identities are highly salient and often explicitly under threat from political news or tweets from political leaders of the out-group.

The Design of Social Media Amplifies the Role of Group-Identity Motivations

Social-media platforms have several features that create a different communication and information-consumption experience than face-to-face interactions (see Bayer, Triệu, & Ellison, 2020), which has important consequences for the form and function of moralized communications. In this section, we examine how certain affordances (see Evans, Pearce, Vitak, & Treem, 2017) of the social-media environment act as an *amplifier* of our group-identity motivations, which ultimately lead to the greater spread of moral-emotional expressions in the context of moralized communications.

Group relationships are highly salient on social-media platforms

In the social-media environment, people interact via information exchange (posting content, sharing content, commenting). In this sense, social media can be viewed as an exaggerated, digital version of a *gossip network*. Gossip is informal information exchange about social events, including the behavior and character of individuals who may not be present (Dunbar & Dunbar, 1998). Face-to-face gossip serves the social function of increasing interpersonal bonding (Dunbar, 2004), which includes preparing for hypothetical situations, interpersonal policing, and advertising one's social value. The social motives underlying the spread of any content on social media may be broadly similar to that of face-to-face gossip networks: We are motivated to exchange information in a way that facilitates positives interactions with those in our social network, or in other words that satisfies our need to belong (Baumeister & Leary, 1995).

We argue, however, that contextual differences between information exchange in face-to-face versus social-media environments create a different locus of social motives. Social motives that drive social-media behavior are entrenched much more in group identity rather than interpersonal relational motives that drive face-to-face gossip. One key contextual difference is group size. Studies have estimated that face-to-face gossip takes place in the context of group sizes of about 150 people, which represents the average number of people one knows personally (Dunbar, 2004). On social media, estimates of the average social network size varies from 350 to 500 (Facebook; A. Smith, 2014) to 700 (Twitter; MacCarthy, 2016). With a substantially larger group size comes a larger audience for which we must maintain a positive status, including those who we may not even know personally. As a result of this larger, less personalized context, our group identities are more salient because a specific group identity is the main relation among our social network rather than an intimate interpersonal relation. Supporting this idea is the large body of work on construal-level theory suggesting that as psychological distance increases, judgment of the self is biased toward high-level, abstract judgments (Trope & Liberman, 2010). For instance, in contexts of greater psychological distance, people are more likely to conform to group norms (Ledgerwood & Callahan, 2012). In other words, to the extent that social-media networks embed us in a larger network that is less familiar, information exchange is more likely to be governed by concerns related to a broader group identity rather than concerns of any one interpersonal relationship compared with face-to-face interactions.

Social-media environments amplify deindividuation

Another notable design feature of computer-mediated communication, including social media, is that people must communicate indirectly with one another through a machine, which necessarily reduces the personal nature of communication and decreases self-awareness (Matheson & Zanna, 1988). A context of reduced self-awareness, particularly in a group-setting such as social media, is ripe for the psychological state of *deindividuation*. Deindividuation refers to a state in which a person experiences reduced self-evaluation in the context of a group, often leading them to behave with less constraints (Diener, Lusk, DeFour, & Flax, 1980; Festinger, Pepitone, & Newcomb, 1952; Postmes & Spears, 1998). More specifically, deindividuation puts a person into a state in which they identify themselves more with the group and conform to group norms more closely (Reicher, 1984), which is a key assumption of the social-identity model of deindividuation effects (Postmes et al., 1998). Insights from this model showcase how the relatively depersonalized nature of social-media communications (design) can promote increased group identification, which comes with a motivation to uphold one's group image.

It is important to note that deindividuation can occur even under conditions in which people are not completely anonymous. For instance, although interactions on Twitter and Facebook are relatively less personal and more anonymous than face-to-face interactions, people often have their pictures and their names on display during communications. However, akin to

people acting in large group crowds in which anyone is technically visible as an individual, the salience of being in a common group (as in the context of communicating to a large social network) can still make people self-categorize more in relation to the group rather than the self. Increased self-categorization in terms of a group identity is the key component that can lead to increased deindividuation (Reicher, 1984) rather than full anonymity per se. The combination of several design features inherent to social media, including less personalized communication, relatively greater anonymity, and salience of being in a large common group (e.g., one's social network might mainly consist of political partisans who hold the same political views) can lead to greater deindividuation and ultimately an enhancement of group-identity motivations that can be fulfilled via the expression of moral emotions.

The feedback delivery system of social media amplifies group-norm conformity

Perhaps one of the most iconic features of social media—the ability to provide immediate and quantifiable social feedback in response to other people's content (e.g., likes, shares, retweets)—may amplify our propensity to express moral emotions in response to morally relevant content (Crockett, 2017). Broadly, the social feedback we receive may make the expression of moral emotions more rewarding. For instance, there is strong evidence that positive social feedback is highly rewarding. People can learn from social rewards (such as smiles and encouragement) just as effectively as material rewards (such as money; Ruff & Fehr, 2014). When people receive positive social feedback, areas of the brain associated with reward such as the striatum are highly active and overlap with brain areas associated with nonsocial reward (Aharon et al., 2004; Izuma, Saito, & Sadato, 2008; Meshi, Morawetz, & Heekeren, 2013; Nitschke et al., 2004; Sherman, Payton, Hernandez, Greenfield, & Dapretto, 2016). These data suggest that signals of positive social feedback are naturally rewarding. Although people receive these signals in normal interaction, they are often ambiguous, rare, or hard to quantify. In contrast, social media allows for unambiguous, ubiquitous, and easy-to-quantify indices of social feedback.

Feedback on social media is not only rewarding in itself but also *social*. Although nonsocial rewards merely indicate that a given behavior is valuable (e.g., food delivery reinforcing lever pressing), social rewards signal that our peers want us to continue the behavior that is being reinforced (Ho, MacGlashan, Littman, & Cushman, 2017). From a very early age, humans have an automatic tendency to infer what others are trying to communicate upon receiving feedback (Bonawitz et al., 2011; Sage & Baldwin, 2011) and use this information to inform future behaviors (Egyed, Király, & Gergely, 2013). For example, when a mother encourages a child to share, the child may infer that sharing is a desired social norm he or she should follow. Likewise, when people receive positive social feedback on social media after expressing outrage about a particular issue, they may automatically infer that expressing outrage about that issue is desired or expected by the group that makes up their social network. The tendency for people to infer the intentions of those providing social feedback leads to *internalization* of the behavior in question (Grusec & Goodnow, 1994). Internalization involves a transition from a behavior being rewarding because it leads to positive feedback to the behavior being rewarding in itself (Grusec & Goodnow, 1994). It is noteworthy that internalization can lead to a behavior being performed in the absence of the feedback that leads to internalization in the first place because they come to view it as normal behavior in their group (Ho et al., 2017). Internalization also leads people to expect the reinforced behavior from others in their social group; that is, they assume it is normative (Vredenburgh, Kushnir, & Casasola, 2015). In the context of outrage expression on social media, the internalization process may make people more likely to express outrage over time, even in the absence of positive social feedback, but also to provide social feedback for others who express outrage about similar issues. This process can create a cascade of outrage amplification within social networks based on conformity to perceived group norms of outrage expression.

Predictions and implications

In this section we examined three design features that can amplify the salience of group identity on social media: large group/audience sizes, less personal interactions, and the social-feedback delivery system. Each of these design components leads to specific predictions and suggests different intervention strategies for reducing out-group-sanctioning moral-emotion expressions that may exacerbate intergroup conflict during social-media communications. Regarding group size, the MAD model predicts that, on average, users embedded in larger social networks (as opposed to smaller social networks), in which users are less familiar with any one user in the network, should show higher group identification and in turn greater moral-emotional expressions when it comes to moralized communications. One important moderating variable for this prediction might be the extent to which a social network demonstrates homophily (e.g., whether users network

with people of a similar political ideology; McPherson, Smith-Lovin, & Cook, 2001). In social networks that are large but mixed in ideology, the predicted effects may change. For example, a user who chooses to connect with large groups of people from across the political spectrum represents someone who might be less prone to perceive threat from political out-groups or might be chronically less identified with their own political group, and therefore less likely to express moral emotions during political conversations.

Considering the effects of deindividuation leads to the prediction that variation in the extent to which communications are personal on different social-media platforms will affect group identification and ultimately moral-emotional expression. One key variable for determining how personal communications are is the extent to which representations of personal identity are salient on the platform. For instance, Facebook is notable for users representing their personal life for others, including posting pictures of themselves, tagging users in photos, and displaying an avatar (profile picture, "story"). In this context people's *individual* identity might be represented on the platform relatively more on the average than platforms such as Twitter. Twitter simply contains one avatar that goes along with people's text or image communications (many of which do not actually contain a picture of the user). Thus, users on Twitter may be more prone to deindividuation, group identification, and in turn more moral-emotional expressions. A recent study comparing emotional expressions on Facebook vs. Twitter supports this prediction (although not in the context of political communications): Social-media users who used Facebook reported that they express their emotion less often than social-media users who use Twitter (Errasti, Amigo, & Villadangos, 2017, p. 1003).

Regarding positive social feedback, the central prediction outlined here is that users who receive more positive social feedback when they express moral emotions should express more of those emotions in the future posts or share more emotional content. This effect might change over time as a result of the process of inferring norms of the network from the social feedback. Users who learned from social feedback of their network that the expression of moral emotions is normative might express specific emotions without sensitivity to feedback because it is a behavior they have internalized.

By focusing on the design features of social media that can amplify the effects of group-identity motivations, the MAD model provides a framework for possible interventions for hostile intergroup communications. If design features of social media can amplify group-identity motivations that at times create barriers for intergroup communications (e.g., when group-identity motivations encourage outrage expression that sanctions the out-group), then design features can also be altered to reduce those motivations. In other words, design features can be altered to reduce the tendency to express out-group-sanctioning emotions by reducing group-identity salience or changing the salient group to a superordinate common group. For instance, platforms could use advertisements or notifications that remind American political partisan users of their common identity as an "American." This could reduce perceptions of out-group threat among Democrats and Republicans. Alternatively, users could be given an option to include a notable icon in their profile that represented their national identity to make salient the common identity in their profiles. Of course, extensive testing would be required to ensure that making one superordinate identity salient for a large group of people did not have unintended consequences of creating new, broader out-group targets.

Supporting these ideas is extensive evidence from lab and field studies showing that common in-group identities improve cooperation (Sherif, 1961), increase out-group empathy (Cikara, Bruneau, Van Bavel, & Saxe, 2014), and reduce implicit bias (van Bavel & Cunningham, 2012). In fact, recent studies on social media found that manipulating shared common identities among rival religious groups reduced expressions of hatred (Siegel & Badaan, 2020). Furthermore, a prediction of the social-identity approach is that people's online behavior will be affected by perceptions of one's in-group, and therefore in-group policing of behavior is likely to be more effective than sanctioning from out-groups. For instance, a field study on Twitter found that criticism via direct messages from the in-group, but not the out-group, reduced people's use of racial slurs in online communications (Munger, 2017). Interventions that harness these aspects of social identity are more likely to improve online discourse.

One important implication regarding the social-feedback delivery system is that the amplification of group-norm conformity can lead to situations in which political partisans appear much more polarized than they are in reality. If people's online expressions are responding to the perceived reinforcement contingencies of their social-media network rather than their own affective state, then this could lead to a false-polarization effect in which members of a network are overperceiving the degree of political polarization (e.g., Levendusky & Malhotra, 2016). However, the MAD model proposes such negative consequences could be altered by changing users' ability to learn from feedback they are receiving. One dramatic change would be to remove the ability for users to quantify how much social feedback

any one post is receiving so that their behavior is not so contingent on social reward. Such changes would represent a notable departure from the current social-media experience but could have a powerful impact on negative experiences such as polarizing political communications.

Summary

In this section, we argued that several design features of the social-media environment including relatively large audience sizes, less personal communication, and the social-feedback delivery system all have the potential to shift people's self-categorization from the individual to the group and amplify group-norm conformity. As discussed above, expressions of moral emotions serve as functional responses to various types of group threats and are likely to increase in contexts in which people evaluate their world in reference to their group identity rather than their personal identity.

Moral-Emotional Content Captures Our Attention

In this section we examine a more basic psychological property pertaining to moralized content that can help explain why it spreads online: Our perceptual systems may be naturally tuned to detect stimuli that are associated with morality and emotion (see Anderson, 2005; Gantman & Van Bavel, 2015). One key feature of social media is that it allows instant access to a massive amount of information. On the one hand, this feature can benefit us by allowing us to learn about or become aware of ideas we would not have otherwise encountered (e.g., news, education, products). However, increased information also comes at a cost. As information access increases, our ability to pay attention to it decreases (H. A. Simon, 1996). Indeed, social media has been described as an "attention economy" (Williams, 2018) because users are bombarded with various types of content that are all competing for our attention. In the typical newsfeed on a social-media platform such as Facebook or Twitter, the average person scrolls through 300 ft (91.44 m) of messages per day (Wade, 2017). In this constant stream of messages, people have milliseconds to scan each message before moving to the next message. Consequently, content that captures our attention more than others has a distinctive advantage in drawing engagement; that is, we must notice content for it to spread online.

We propose that social-media messages containing moral-emotional expression may be shared more than other types of messages in part because moral and emotional content both have the ability to capture our attention more than other types of content. "Attention" refers to the selective processing of information while ignoring other information (Dijksterhuis & Aarts, 2010). Because our perceptual systems are constantly bombarded with sensory information, higher cognitive processes can use only a small amount of it. Thus, by "greater attentional capture," here we mean prioritized, selective visual processing, including (a) rapid and automatic processing and/or (b) shifting of cognitive resources to the attended stimuli over others (see, e.g., Öhman & Mineka, 2001). In this way, moral-emotional content may be prioritized relative to other content and therefore has the ability to draw increased engagement and spread further in online networks.

Moral and emotional content may be particularly prone to capturing our attention because it is *motivationally relevant*. A stimulus is motivationally relevant if it can affect an active or ongoing goal, and stimuli that affect goals tend to be prioritized in visual attention (Dijksterhuis & Aarts, 2010). Below we outline evidence that both moral and emotional content capture attention more than other content that is less motivationally relevant.

Morality and attention

Moral content is motivationally relevant because morality is associated with numerous social motivations, including needs related to control over our world (Kay, Gaucher, McGregor, & Nash, 2010), social justice (Lerner & Miller, 1978), and belonging in groups (Baumeister & Leary, 1995; Haidt, 2012). More broadly, morally relevant stimuli often provide key social information relevant to our well-being, such as information about people or groups, that act in ways that could help or harm us (e.g., cheating, stealing, giving; Cosmides & Tooby, 1992; Fiske, Cuddy, & Glick, 2007). Thus, morality is likely salient to most social-media users on a regular basis.

Because morality is motivationally relevant, it is not surprising that our cognition may be naturally tuned to detect morally relevant stimuli. For instance, research on impression formation has consistently found that signs of bad behavior immediately capture our attention when forming character judgments (Fiske, 1980; Pratto & John, 1991; Skowronski & Carlston, 1989) and increase rates of learning about the traits of others (Siegel, Mathys, Rutledge, & Crockett, 2018). Faces are also attended to more when they are paired with negative morally relevant information (Anderson, Siegel, Bliss-Moreau, & Barrett, 2011). This attention-capturing capacity also translates to representations of morality in language and images. Moral words presented near the threshold for conscious awareness appear to "pop

out" in visual experience compared with neutral words (Gantman & Van Bavel, 2014, 2016; but see Firestone & Scholl, 2015). Furthermore, extensive research has found that people are sensitive to justice and other moral concerns (Lerner & Miller, 1978; Schmitt, Baumert, Gollwitzer, & Maes, 2010), and violations of justice lead to a motivation to restore justice and make moral content more salient in the environment (Hafer, 2000; Kay & Jost, 2003). The role of attention is especially important given recent evidence that increased attention to a decision option is linked to judgments of wrongness, blameworthiness, and even legal-punishment decisions (Granot, Balcetis, Schneider, & Tyler, 2014; Pärnamets et al., 2015). Together, these data suggest that morally relevant content is more likely to capture attention over more neutral content and facilitate moral contagion on social media. Recent findings directly support this claim: Moral and emotional words that captured more attention in a laboratory setting were associated with greater sharing on social media when they appeared in messages during political communications (Brady, Gantman, & Van Bavel, 2020; see Fig. 5).

Emotion and attention

Emotional stimuli are motivationally relevant because they typically threaten or promote well-being and thus require immediate response (e.g., detection of snake-like objects in a field; Ohman, Flykt, & Esteves, 2001). Emotional stimuli are usually motivationally relevant in social settings because they help determine how to navigate social interactions (Campos, Mumme, Kermoian, & Campos, 1994). The motivational relevance of emotional stimuli may be most obvious when it comes to real-life objects such as snakes or people, but the human brain can also assess motivational relevance in content that represents emotion, such as language (Kissler, Herbert, Winkler, & Junghofer, 2009). The emotional significance of language is extracted from the brain rapidly within 250 ms (Kissler et al., 2009), and possibly even prelexically within 100 ms (Bernat, Bunce, & Shevrin, 2001), suggesting that emotional language on social media could draw in our attention immediately.

Multiple studies demonstrate that emotional stimuli spontaneously capture attention during undirected viewing (Chen, Shechter, & Chaiken, 1996; Kissler, Herbert, Peyk, & Junghofer, 2007; Ortigue et al., 2004; Skrandies, 1998). In the social-media environment, however, multiple forms of content are specifically developed and selected for their ability to capture our attention (Rose-Stockwell, 2017), possibly creating an environment in which attention must be captured under conditions of limited cognitive resources. A large body of research suggests that emotional language draws attention more than other types of words, even under such conditions (Anderson, 2005; Anderson & Phelps, 2001; Keil & Ihssen, 2004; Milders, Sahraie, Logan, & Donnellon, 2006). Emotional stimuli can also draw attention away from ongoing visual goals (Arnell, Killman, & Fijavz, 2007; Ciesielski, Armstrong, Zald, & Olatunji, 2010; Most, Smith, Cooter, Levy, & Zald, 2007; Most & Wang, 2011). Both in conditions of normal viewing and limited cognitive resources, emotional stimuli, including emotional representations in language, naturally capture our attention more than neutral stimuli.

Predictions and implications

The basic prediction of the MAD model is that moral and emotional content capture more attention than other types of more neutral content when users are interacting with their social-media feeds. Thus far, evidence from a study investigating a small but tightly controlled set of moral and emotional stimuli supports the prediction: Moral and emotional words captured more attention than neutral words, and their attentional-capture capacity was associated with sharing during political communications on Twitter (Brady, Gantman, & Van Bavel, 2020).

One key moderating variable for the prediction that moral and emotional content draw engagement by capturing attention is what counts as moral content from the perspective of the individual users or their social networks. For instance, a large body of work suggests that American liberals and conservatives base their sense of morality on different values (Graham, Haidt, & Nosek, 2009; see also Kugler, Jost, & Noorbaloochi, 2014). As a result, the specific content that is construed to be in the domain of morality—and therefore interpreted as socially and motivationally relevant—will vary depending on the political ideology of the user. Recent data support this idea: In a sample of more than 11 million tweets from 25,000 American Twitter users, liberals were more likely to express their morality in terms of fairness concerns, whereas conservatives were more likely to express their morality in terms of loyalty, authority, and purity concerns (Sterling & Jost, 2018).

More specifically, group identity should interact with attentional capture: A social network that is formed on the basis of concerns about a specific moral issue will show amplified attentional capture for any content referencing that specific issue. For instance, users embedded in social networks that are composed of antivaccine proponents (who consider vaccinations morally wrong because of their supposed harm) are especially drawn to content that is related to the effects of vaccines and specifically content that supported their moral

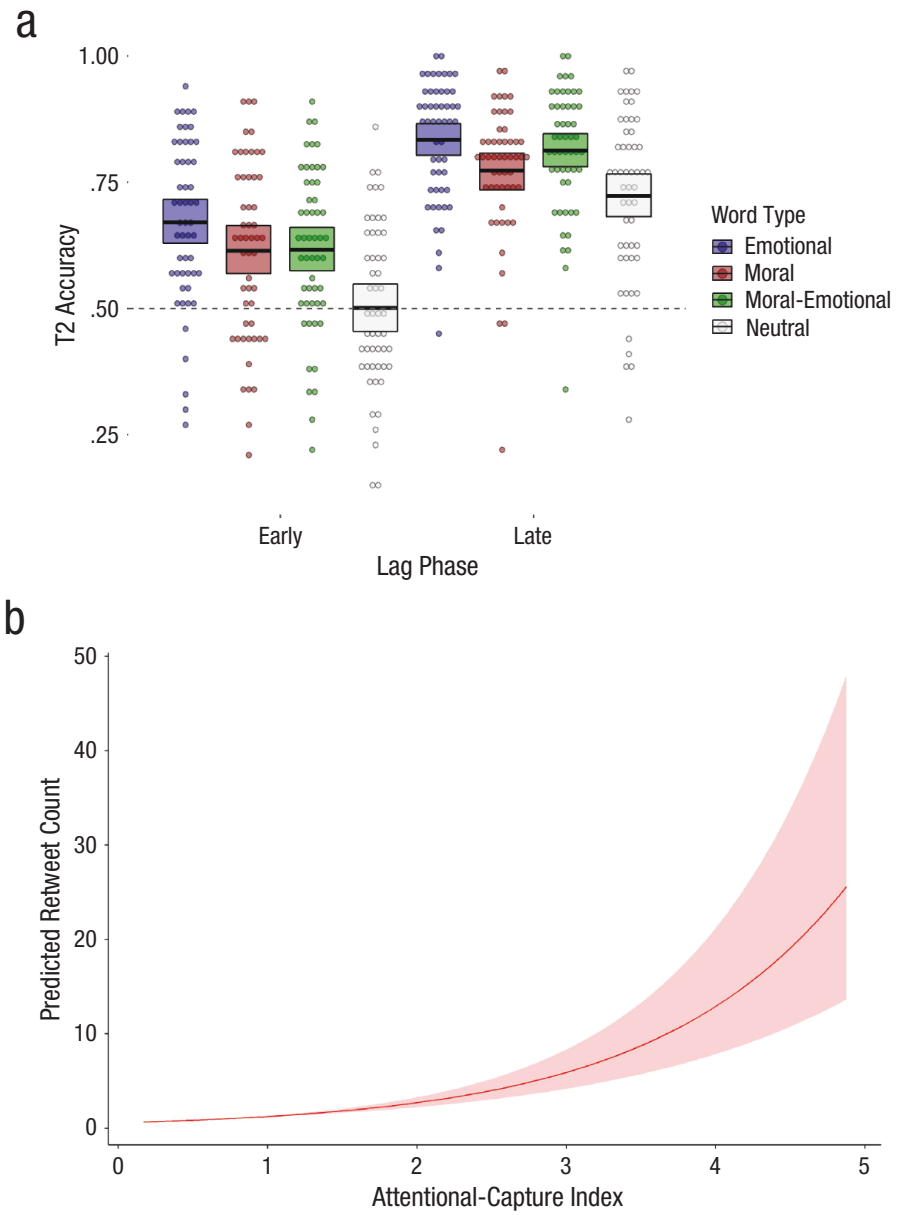

Fig. 5. Association of online sharing with the ability for moral and emotional content to capture attention. Moral and emotional words garnered greater attentional capture as measured by accuracy in an attentional blink paradigm in a laboratory study. In the box plot in (a), accuracy at Time 2 (T2) is graphed as a function of lag phase, separately by word type. The line in the middle of each box represents the mean, and the top and bottom edges of the box represent +1 *SEM* and –1 *SEM*, respectively. In (b), predicted tweet count is graphed as a function of attentional-capture index. The shaded area around the curve represents the 95% confidence interval. Tweets with a greater attention-capture value as assessed by specific words in the tweet were associated with greater expected retweet counts. The attentional-capture index was calculated on the basis of the mean attentional-capture data from our lab study for each word present in a tweet. From Brady, W. J., Gantman, A. P., & Van Bavel, J. J. (2020). Attentional capture helps explain why moral and emotional content go viral. *Journal of Experimental Psychology: General, 149*, 746–756. Copyright © 2020 American Psychological Association. Adapted with permission.

views about vaccines (Schmidt, Zollo, Scala, Betsch, & Quattrociocchi, 2018). Ultimately, to understand the explanatory role of attention regarding social-media engagement, it is necessary to identify the specific moral values that are most salient in the network, which should be a direct function of the dominant group

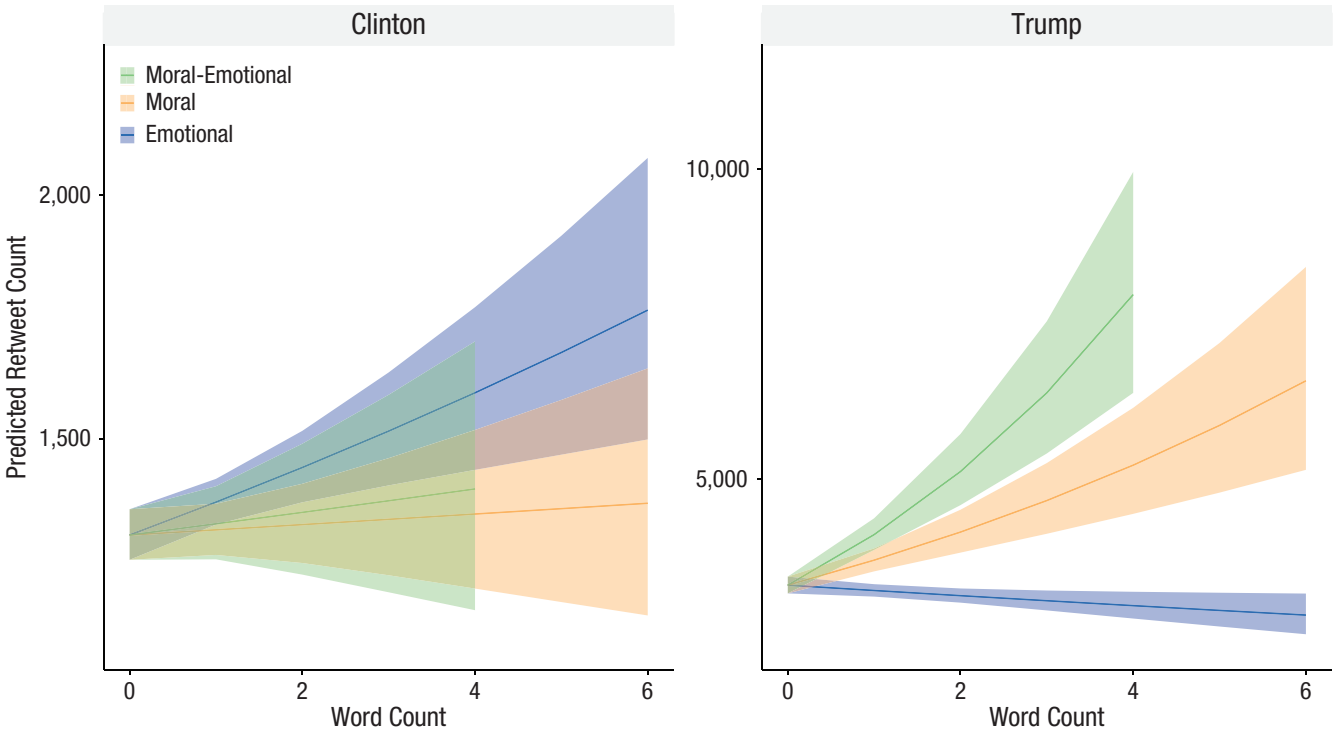

Fig. 6. Moral-emotional language use by Donald Trump and Hillary Clinton during the 2016 U.S. presidential campaign. Predicted retweet count is graphed as a function of word count, separately by word type. The shaded areas around the curves represent 95% confidence intervals. Donald Trump's use of moral-emotional language was significantly associated with increased diffusion on Twitter (27% expected increase in retweets for every moral-emotional word added to a tweet), whereas Hillary Clinton's was not (2% expected increase). From Brady, W. J., Wills, J. A., Burkart, D., Jost, J. T., & Van Bavel, J. J. (2019). An ideological asymmetry in the diffusion of moralized content on social media among political leaders. *Journal of Experimental Psychology: General*, *148*, 1802–1813. Copyright © 2019 American Psychological Association. Adapted with permission.

identity in a network. Indeed, extensive work suggests that social identity drives a wide range of perceptual processing, including selective attention to identity-relevant stimuli (for a review, see Xiao, Coppin, & Van Bavel, 2016).

Another moderating variable may be the extent to which a user has attended to the content previously. Attentional capture can be reduced as a result of previous attentional engagement to a stimulus (Klein, 2000), and this can lead to habituation or "inhibition" effects for moral and emotional stimuli that generally capture attention. For instance, one news site that rose to fame by continually producing attention-grabbing clickbait news—*Upworthy*—recently began losing millions of users, leading it to change its business model away from attention-grabbing clickbait (Sutton, 2016). It may become easier for people to ignore content that they engaged with previously.

The predictions presented in this section have several important implications for information consumption on social media that pertains to moral values (e.g., politics, activism, disinformation). If moral and emotional content has an advantage in the attention economy of social media, then moral-emotional expression may be leveraged by social-awareness campaigns, political groups (including disinformation campaigns), and businesses as an efficient method of drawing greater engagement among competing content. For instance, cases of viral prosocial online campaigns that raised more than $100 million (e.g., the ALS Ice Bucket challenge, Save Darfur campaign) specifically appealed to people's sense of morality and utilized emotional appeals (Van Der Linden, 2017). Prosocial campaigns need to get noticed to increase their donations, and targeted moral-emotional appeals can help explain how they do.

Likewise, political campaigns can leverage moral contagion to draw attention to their ideas and policies, which appeared to be effective for most presidential candidates and members of congress during the 2016 U.S. presidential election (Brady, Wills, et al., 2019; see Fig. 6). In some cases, the use of evocative moral and emotional content can draw attention that is beneficial for political outcomes regardless of how much political division is created from that attention. For example, many argued that Donald Trump's use of evocative content led to increased attention from media across the political spectrum and that this exposure ultimately

helped him win the election (e.g., Sillito, 2016). Indeed, we have found not only that Donald Trump's tweets containing moral-emotional language were far more likely to go viral than Hillary Clinton's (Brady, Wills, et al., 2019) but also that the effect size for Trump was larger than the average person discussing "hot-button" political issues (Brady et al., 2017). More research is required to determine how this varies for liberals versus conservatives and in what contexts increased attention might lead to unpopularity or negative consequences, as in the case of insensitive moral statements that lead to viral firestorms against the expresser (Pfeffer et al., 2014) or empathy for people who transgress social norms—known as the paradox of viral outrage (Sawaoka & Monin, 2018).

On the other hand, moral and emotional appeals that capture attention can be exploited by disinformation profiteers, as in the case of fake news spread around the 2016 U.S. election that was more likely to be emotional and novel (Vosoughi et al., 2018). However, this also suggests that attention can be used as a way to combat attraction to fake news. For instance, shifting people's attention to other aspects of fake news content such as the trustworthiness of the source may also help to combat the consumption and spread of fake news (Pennycook & Rand, 2019).

Finally, the attention-grabbing nature of moral and emotional content also has important implications for the rise of "psychographic marketing," which attempts to leverage psychological profiles of individuals as a marketing strategy (e.g., Dutta-Bergman, 2004). By understanding the type of content that is motivationally relevant for different groups with differing moral values, companies can appeal to these moral values to better draw attention to products and services during social-media marketing. This may be particularly valuable for brands targeting the "conscious consumer" (Loureiro & Lotade, 2005), where the viral spread of moralized content could shift attitudes positively toward their brand.

Summary

In this section, we argued that moral and emotional content are prioritized in visual attention because such content is motivationally relevant from both a biological and social standpoint. The attention-capturing properties of moral and emotional content can give it an advantage in the attentional economy of social media, in which content must break through the immense noise of other items in our personal-content feeds. Whereas moral and emotional content have the capacity to capture our attention broadly speaking, what counts as "moral content" will depend on the specific values held by the individual or what is normative for the social network as a whole. The impact of attention on the spread of moralized content therefore depends jointly on the stimuli themselves and the values inherent to one's group identity.

The Design of Social Media Amplifies Attention to Moral-Emotional Content

Social-media platforms are specifically designed with the goal of keeping users' attention sustained on the platforms (Williams, 2018). Sustained attention leads to greater engagement, which leads to greater profits for the social-media companies (Alter, 2017). Thus, design features such as content algorithms and notifications that remind us of activity on the platform can amplify our attention to content that we are more likely to notice in the first place.

Content algorithms act as an external attentional filter

Content algorithms expose users to information via their content feed that is more likely to draw their engagement on the basis of many variables, including their previous behavior, predictions about what they will enjoy or care about, and other unknown variables that are consistently changing (Agrawal, 2016). Ultimately, content algorithms are designed to increase engagement and profit for the platforms (Rose-Stockwell, 2017; see Fig. 7). In this way, content algorithms that are vital to social-media platforms act as an external "attentional filter" by preselecting moral or emotional content that our perceptual system has a tendency to notice in the first place.

For example, according to software engineers at YouTube, their algorithm learned that the best way to get people to watch more videos was to show people videos loaded with speculation about popular events (Popken, 2018). There is some evidence that this algorithm can deliver increasingly extreme video content (Chaslot, 2018)—precisely the type of content that would be expected to generate moral-emotional reactions. However, it is also important to consider the role of politically extreme communities in placing extreme content online in the first place (Munger & Phillips, 2019). Social-media algorithms act as a significant filter that can increase the chances of some content drawing social feedback over others, and the interaction of content algorithms and people's natural tendencies must be considered in explaining how social feedback shapes people's online behavior.

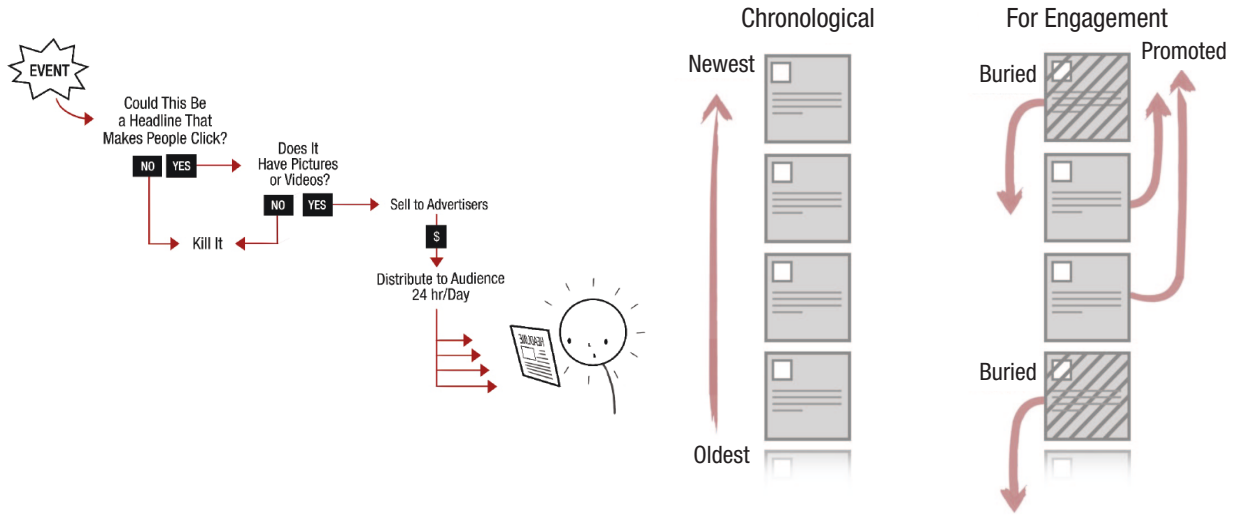

Fig. 7. User engagement is an interaction of attraction to content and selective exposure based on algorithms. The positive social feedback received on social media is a combination of people's natural attraction to content and selective exposure to content based on social-media algorithms that are designed to boost engagement and ultimately profit for advertisers. If people's online behavior is shaped by social learning from feedback received through likes and shares, this learning is partially governed by content that algorithms make the most available for people to interact with. Images created by and reproduced with the permission of Tobias Rose-Stockwell.

Notifications amplify sustained attention

Notifications that inform users about their account activity is now ubiquitous across social-media platforms. Notifications update users when new social interactions occur or when new informational content is available for them to view. These notifications have transformed the social-media experience. People's attention is demanded and directed to specific content by the platform. For instance, recent research estimates that most people attend to phone notifications within a few minutes of the notification, and the rate of people who attend within minutes is notably higher when the notification is based on recent social interactions (e.g., a social-media chat message; Chang & Tang, 2015). Notifications even direct our attention when it is meant to be directed elsewhere in current activities not related to the phone (Shirazi et al., 2014). Simply put, notifications can amplify our attention to content that is morally relevant to our specific values or emotional in nature, in particularly if we have already interacted with the content and the notifications pertain to our previous interaction. For instance, if we share a news story showcasing morally offensive behavior from a political outgroup and also express outrage as part of our post, we may get notifications that other people liked/shared our post or provided their own moralized comments to the content we shared. Just as the social-feedback delivery system can amplify moral contagion as argued above, notifications can sustain our attention to content that is naturally noticed by our perceptual systems.

Predictions and implications

Predictions derived from this section are tied into design interventions that could change the content we are likely to interact with on social media. One key prediction is that manipulating content algorithms to reduce the amount of moralized content pushed to users' feeds will reduce the rate of posting and sharing it, even given that people are naturally attuned to such content compared with more neutral content. This might have the important implication of reducing the impact of disinformation campaigns that draw on moral and emotional content to provoke intergroup conflict. The dilemma for social-media companies is that this might conflict with pushing content that draws the most engagement—and ultimately the most profit. On the other hand, users on the platform can exploit algorithms no matter how often they are changed to produce content containing features most likely to be promoted by the content algorithm. Thus, the impact of content algorithms on posting and sharing of moral and emotional content is a dynamic process that also requires understanding the motivations people may have for exploiting the social-media design to draw engagement.

There are also other options in addition to manipulating the actual content algorithm for controlling the flow of moral and emotional content that users are exposed to in their personal-content feeds. For instance, Twitter announced recently that it will be banning all political advertisements from its platform to prevent

disinformation profiteers from leveraging influence in the 2020 U.S. election (Conger, 2019). This decision has the potential to manipulate the flow of moral and emotional content in people's feeds, even though the content algorithms are not necessarily changed.

Regarding notifications, we predict that a notification system that reduces sustained attention to social-media platforms can reduce the amplification of toxic moral and emotional content that targets political out-groups. For instance, people may be more likely to respond in a productive manner if they are involved in a political debate in a social-media thread and they are not constantly being notified of immediate responses in the debate. By reducing attention to the thread, people might be able to distance themselves from the conversation and be less emotionally aroused during the next response. Although studies have not investigated this specific idea, recent work found that "batching" notifications so users only received notifications three times per day reduced stress and increased well-being (Fitz et al., 2019).

Summary

In this section, we examined how content algorithms and notifications—two design features inherent to social media—can amplify our natural attraction to moral and emotional content. Content algorithms act as an exogenous attention filter that preselects content we are most likely to engage with. Notifications sustain our attention to social-media content by directing it back to the platform, even when we are doing other tasks. As these features keep our attention to specific content we previously interacted with, they can serve to guarantee extended attention to moral and emotional content we may have noticed and engaged with on the platform previously.

The Design of Social Media Facilitates the Spread of Emotional Expressions

In previous sections, we examined how design features of social media specifically amplifies social (group-identity motivations) and cognitive (attentional capture) psychological tendencies. In this section, we focus on design features that facilitate moral contagion in the sense that they broadly make emotional expressions more likely to spread compared with other media. These features include how emotions are represented in the social-media communication medium, the available options for people to react to content, and the lowered cost of expressing emotions that would be costly during face-to-face interactions. Each of these features amplifies the ability of moral-emotional expressions to spread and can help to explain the prevalence of moral contagion on social media.

Symbolic representation of emotion

One notable feature of the social-media communication medium (and other computer-mediated environments) is that emotion expressions occur via symbolic representations in language and images. Symbolic representations of emotion expression have at least three properties that differ from nonverbal behavioral expression typical of face-to-face interaction (e.g., facial expressions), and each of these properties can affect how the emotions spread to others (see Peters & Kashima, 2015). We discuss these properties and how they might relate to moral contagion.

First, emotion expression online is more *static* than nonverbal expression. Once expressers post messages labeling their emotions, that expression stays the same over time so long as the message remains (it could be deleted or replaced with a new message but it cannot be altered on many platforms), unlike fleeting facial expressions or other nonverbal behaviors. Screen captures can also ensure the expression is static, even if the original post is deleted. Static expressions of emotion such as those inherent to social media may spread to a larger number of people and for a longer amount of time because the expression is available to be perceived for as long as it remains online.

Not only might the static nature of emotional expressions give them a greater likelihood of being discovered, but also, insofar as people are aware of this static nature, it may alter what they share or post. If people are aware that the content they share or post may exist forever, they could be motivated to post content that is more "universal" or "objective" in nature. Because universalism and objectivity is a core feature of moral beliefs and attitudes (Goodwin & Darley, 2012; Singer, 1961; Skitka et al., 2005; van Bavel et al., 2012), one possibility is that people are either explicitly or implicitly more likely to post content expressing strong moral values as a result of an awareness of greater longevity of their posts. In other words, insofar as we want our content to stand the test of time, we may assume that the expression of moral emotions and moral content more generally could be well equipped to fit this goal.

Second, the process of sharing emotion expressions online maintains *higher fidelity* than the spread of nonverbal behavioral expression. When a perceiver reads and shares an expresser's content, the sharing fully reproduces the original content, allowing anyone who perceives the shared content to glean identical information as the original perceiver's expression. On the one hand, this feature makes it more likely for someone to

understand a moral violation that is the object of an emotion expression, which allows the content to affect and possibly motivate someone to share the content even if they are far removed from the original experiencer in terms of network positioning. On the other hand, the content that is shared is only high fidelity in the sense that it copies the way the original poster *described* the experience. In other words, social media allows for high-fidelity sharing of emotion expression, but the original emotion expression could be an incomplete representation of the poster's experience (i.e., devoid of contextual information left out by the poster). In fact, the message could be taken completely out of context in a way that changes the meaning and allows for a stronger moral-emotional claim by people who wish to share it. This could heighten emotional discord if the original expression is misconstrued.

Third, through language and images, emotion expression online has *object representation* that is relatively lacking in nonverbal cues. When online, one can represent the specific stimulus that caused the emotion through language or images (e.g., a detailed description or meme of the morally offensive actions of a political out-group member), but nonverbal behaviors (e.g., a smile) typically do not directly represent the eliciting object. Object representation gives the original emotion a greater chance of spreading to others because people can perceive the details of the eliciting conditions even though they were not there. With some content such as images and videos, the object of the emotion might even be represented as if anyone viewing the content was actually present in the eliciting situation. However, the object representation is constrained in unique ways on each social-media platform, and this may have important consequences for moral contagion. For instance, Twitter recently changed its platform to allow users to post longer messages (from 140 to 280 characters). An initial analysis suggests that this subtle change in the design actually influenced people's tendency to post content that was relatively more analytical and also more polite (Jaidka, Zhou, & Lelkes, 2019). In this way, the subtle design features of each platform will elicit different patterns of expression and behavior.

Rapid response options

Another key design feature of social media that may facilitate the spread of moral-emotional content are response options that encourage quick, rapid responses: Users can like, share, or retweet all in the amount of time it takes to blink (Crockett, 2017). There is indirect evidence that such features encouraging fast responding may increase the dissemination of moral and emotional content. For example, moral decisions driven by emotional reactions are associated with faster responding (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Greene et al., 2004), and when people are forced to make moral decisions quickly they tend to rely on emotional reactions (Suter & Hertwig, 2011). Likewise, when people think about their attitudes in moral terms they are faster to respond (van Bavel et al., 2012). This is in contrast to the type of moral reasoning that often requires greater deliberation and draws on different principles (Greene, 2008; Kohlberg, Levine, & Hewer, 1983). For instance, people may become more parochial in their punishment decisions when they are responding swiftly to moral transgressions and more just when they are using deliberation (Yudkin, Rothmund, Twardawski, Thalla, & Van Bavel, 2016). If people are more likely to rely on emotional reactions for moral decisions that are made rapidly, then design features encouraging rapid reactions in the context of moral and political discussion are likely to facilitate the spread of moral-emotional content (i.e., moral contagion).

Reduced personal costs of emotion expression

Another important feature of social-media communication is that information is exchanged in a less personal, digital environment that reduces the costs of using information exchange for intra- and intergroup policing compared with offline interactions (e.g., face-to-face gossip). On social media compared with other contexts, people are more likely to experience moral outrage (Crockett, 2017)—an emotion that is associated with punishment behavior (e.g., Fehr & Fischbacher, 2004). In contrast, one study found that interpersonal policing accounted for less than 5% of total gossip content during face-to-face communications (Dunbar, 2004). One reason may be that the personal costs—in terms of the possibility of retaliation and empathetic distress—of policing in-group and out-group members on social media are highly reduced (Crockett, 2017). With group identities more salient and reduced personal costs, people on social media are much more likely to use information exchange to derogate out-groups via the expression of moral emotions such as outrage.

Predictions and implications

In this section we examined design features of the social-media communication environment that are likely to amplify the spread of emotional content on social media. One key prediction from the MAD model is that moral-emotional expressions can spread faster and further in social-media networks compared with other media. Other predictions derived from this

section pertain to specific interventions based on design decisions that could potentially reduce people's reliance on emotional reactions during social-media communications. More specifically, the MAD model predicts that the following design changes could result in lowered reliance on emotional reactions: (a) diminishing the fidelity of emotional representations each time they are shared, (b) making reactions take longer or require some sort of cognitive reflection, or (c) making personal effects of targeting someone with an emotional response more salient.

One example of such a design change could be if users were met with an "empathetic prompt" reminding them that what they are about to say is potentially hurtful (Rose-Stockwell, 2018). Such interventions could potentially reduce negative intergroup or interpersonal communications by forcing people to take a little more time to reflect on their messages. Another potential design change is to make available sections of social-media sites that have a salient goal of providing social support to users. Indeed, existing websites with these norms have been shown to improve mental health (Zaki, 2019).

Recognition that subtle design features—which were designed with the goal to increase profits—can have a notable impact on the spread of moral-emotional content has many important implications for how we envision social technologies and their impact on interpersonal and intergroup interactions in the future. A narrow focus on designing social-media environments to increase engagement can lead to negative unintended consequences for individuals and the organizations promoting it. For instance, the same design features on Twitter that allowed people to organize protests to support the growth of democracy in authoritarian regimes (McGarty, Thomas, Lala, Smith, & Bliuc, 2014) also allowed misinformation and conspiracy theories to flourish (Lazer et al., 2018). More broadly, considering how design features affect the spread of emotional expressions is important because the spread of emotions can have the same impact on social behaviors as they do when expressed offline because they achieve a communicative function (Van Kleef, 2009, 2017). We believe that a thoughtful analysis of these design features not only is critical for understanding the psychology of user behavior online but also provides a fertile—and critically important—area for future research.

Considering how design features of social media may influence our emotion expression and experience may be crucial for studying morality and emotion as humans move more fully into the digital age. In particular, the design features of social media may affect the functions of emotions in ways current emotion theories may not be well equipped to explain. For instance, recent work reviewing research on behavior in offline contexts suggested that moral outrage can have "upsides" such as motivating collective action (Spring, Cameron, & Cikara, 2018). Although this is certainly true in some contexts, the design features of social media might create a context that severely limits some of the upsides of outrage (Brady & Crockett, 2018). Although the question of whether moral emotions such as outrage are good or bad for society is ultimately a philosophical question (Nussbaum, 2016; Srinivasan, 2018), we welcome future research to determine where they promote the goals of individuals and groups and where they might undercut their goals or lead to aversive downstream consequences (e.g., by creating unintended false polarization).

Furthermore, the various ways social-media design features can affect our emotion experience and expression has implications for theories of emotion regulation. Emotion-regulation theories have long posited that a key reason we regulate our emotions is to maintain appropriate responses for changing environments (Gross, 1999). However, the idea that the environment itself can regulate our emotions, in the sense of the environment having goals that influence our emotional states, has received little attention. For example, if specific features of social media are created to increase engagement and ultimately revenue rather than to specifically enhance human well-being (Rose-Stockwell, 2017), then in a very real sense social-media platforms can regulate our emotions in ways we might not be aware of or in ways that are not aligned with our personal goals. Future research is required to determine how intra- and interpersonal emotion-regulation goals of people interact with the design goals of digital environments in ways that influence our emotions. Such research could help to inform software engineers and organizations, allowing them to use a more psychologically informed approach to social-media design.

Summary

In this section we examined how social media constrains emotion expression to symbolic representations in language and images, which facilitates the spread of emotion because the expressions are static and high fidelity and represent the eliciting object of the emotion. Furthermore, we argued that quick response options and the reduction of personal costs make it more likely for people to post or share emotional responses.

Conclusion

Social-media usage is still growing by hundreds of millions of users every year (Statistica, 2018), and platforms

such as Twitter, Facebook, and YouTube have become the dominant public space to learn about and discuss morality and politics (Duggan & Smith, 2016). As digital interactions become one of the most common social contexts, it is increasingly important for scientists to understand why people behave as they do online and what consequences the shift from offline to online communication contexts has for our daily lives. Here, we propose a model that helps to explain why moralized content spreads. This is more important than ever before to understand because of its growing implication in our everyday moral and political life. This includes societal-level phenomena such as political uprisings, national elections, hate speech, violence, political polarization, and even international conflict.

The MAD model integrates theories of intergroup interaction from social psychology with theories of information processing from cognitive psychology to help situate the phenomenon of moral contagion in online networks as a natural extension of existing psychological tendencies. Indeed, our brains may be hardwired to identify with social groups (Cosmides, Tooby, & Kurzban, 2003) and attend to information that is motivationally relevant for our social and biological survival. These tendencies may play out on social media similarly to how they are implicated in other communication contexts, whether in face-to-face communications or other digital media.

However, a key contribution of the MAD model is to underscore how our natural psychological tendencies are amplified by the specific design features present in the social-media environment. Specifically, the MAD model highlights how social media can amplify group identification and group-based emotions, enhance our attention to moral and emotional content, and increase the ability for emotions to spread further and more quickly than in other contexts. More broadly, the model calls attention to the fact that social-media platforms are not "neutral" in the sense that small decisions that constrain human behavior can lead to societal-level consequences pertaining to how humans relate to one another. Understanding this interaction between human moral psychology and social media is urgently needed and relevant to a number of issues. We hope that the framework and predictions presented here can help to explain and spark future research on civic engagement and activism, political polarization, propaganda and disinformation, and moralized consumer behavior, as humans become more immersed than ever before in digital social technologies.

Transparency

Action Editor: Jennifer Wiley
Editor: Laura A. King

Declaration of Conflicting Interests
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding
This work was supported by National Science Foundation Award 1808868, Democracy Fund Grant R-201809-03031, and the New York University Research Challenge Fund.

ORCID iDs

William J. Brady https://orcid.org/0000-0001-6075-5446
Jay J. Van Bavel https://orcid.org/0000-0002-2520-0442

References

Abrams, D., & Hogg, M. A. (2004). Metatheory: Lessons from social identity research. *Personality and Social Psychology Review*, *8*, 98–106.

Agrawal, A. (2016, April 20). What do social media algorithms mean for you? *Forbes*. Retrieved from https://www.forbes.com/sites/ajagrawal/2016/04/20/what-do-social-media-algorithms-mean-for-you/#3c9ffaffa515

Aharon, I., Etcoff, N., Ariely, D., Chabris, C. F., O'Connor, E., & Breiter, H. C. (2004). Beautiful faces have variable reward value. *Neuron*, *32*, 537–551.

Algoe, S. B., Haidt, J., & Gable, S. L. (2008). Beyond reciprocity: Gratitude and relationships in everyday life. *Emotion*, *8*, 425–429.

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*, 211–236.

Alter, A. (2017). *Irresistible: The rise of addictive technology and the business of keeping us hooked*. New York, NY: Penguin Press.

Anderson, A. K. (2005). Affective influences on the attentional dynamics supporting awareness. *Journal of Experimental Psychology: General*, *134*, 258–281.

Anderson, A. K., & Phelps, E. A. (2001). Lesions of the human amygdala impair enhanced perception of emotionally salient events. *Nature*, *411*, 305–309.

Anderson, E., Siegel, E. H., Bliss-Moreau, E., & Barrett, L. F. (2011). The visual impact of gossip. *Science*, *332*, 1446–1448.

Aquino, K., & Reed, A. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, *83*, 1423–1440.

Arnell, K. M., Killman, K. V., & Fijavz, D. (2007). Blinded by emotion: Target misses follow attention capture by arousing distractors in RSVP. *Emotion*, *7*, 465–477.

Arnold, M. B. (1960). *Emotion and personality: Vol. 1. Psychological aspects*. New York, NY: Columbia University Press.

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. B. F., . . . Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences, USA*, *115*, 9216–9221.

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

Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The role of social networks in information diffusion. In *Proceedings of the 21st International Conference on World Wide Web* (pp. 519–528). New York, NY: Association for Computing Machinery. doi:10.1145/2187836.2187907

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, *26*, 1531–1542.

Barrett, L. F. (2013). Psychological construction: The Darwinian approach to the science of emotion. *Emotion Review*, *5*, 379–389.

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, *117*, 497–529.

Bayer, J. B., Triệu, P., & Ellison, N. B. (2020). Social media elements, ecologies, and effects. *Annual Review of Psychology*, *71*, 471–497.

Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, *49*, 192–205.

Bernat, E., Bunce, S., & Shevrin, H. (2001). Event-related brain potentials differentiate positive and negative mood adjectives during both supraliminal and subliminal visual processing. *International Journal of Psychophysiology*, *42*, 11–34.

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*, 322–330.

Brady, W. J., & Crockett, M. J. (2018). How effective is online outrage? *Trends in Cognitive Sciences*, *23*, 79–80.

Brady, W. J., Gantman, A. P., & Van Bavel, J. J. (2020). Attentional capture helps explain why moral and emotional content go viral. *Journal of Experimental Psychology: General*, *149*, 746–756.

Brady, W. J., Wills, J. A., Burkart, D., Jost, J. T., & Van Bavel, J. J. (2019). An ideological asymmetry in the diffusion of moralized content on social media among political leaders. *Journal of Experimental Psychology: General*, *148*, 1802–1813.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences, USA*, *114*, 7313–7318.

Branscombe, N. R., Ellemers, N., Spears, R., & Doosje, B. (1999). The context and content of social identity threat. In N. Ellemers & R. Spears (Eds.), *Social identity: Context, commitment, content* (pp. 35–59). Oxford, England: Blackwell Science.

Brewer, M. B. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin*, *86*, 307–324.

Brown, R., González, R., Zagefka, H., Manzi, J., & Ćehajić, S. (2008). Nuestra culpa: Collective guilt and shame as predictors of reparation for historical wrongdoing. *Journal of Personality and Social Psychology*, *94*, 75–90.

Campos, J. J., Mumme, D. L., Kermoian, R., & Campos, R. G. (1994). A functionalist perspective on the nature of emotion. *Monographs of the Society for Research in Child Development*, *59*, 284–303.

Chang, Y. J., & Tang, J. C. (2015). Investigating mobile users' ringer mode usage and attentiveness and responsiveness to communication. In *MobileHCI '15: Proceedings of the 17th International Conference on Human-Computer Interaction With Mobile Devices and Services Adjunct* (pp. 6–15). New York, NY: Association for Computing Machinery. doi:10.1145/2785830.2785852

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception-behaviour link and social interaction. *Journal of Personality and Social Psychology*, *76*, 893–910.

Chaslot, G. (2018, February 1). How algorithms can learn to discredit the media. *Medium*. Retrieved from https://medium.com/@guillaumechaslot/how-algorithms-can-learn-to-discredit-the-media-d1360157c4fa

Chen, S., Shechter, D., & Chaiken, S. (1996). Getting at the truth or getting along: Accuracy- versus impression-motivated heuristic and systematic processing. *Journal of Personality and Social Psychology*, *71*, 262–275.

Christophe, V., & Rimé, B. (1997). Exposure to the social sharing of emotion: Emotional impact, listener responses and secondary social sharing. *European Journal of Social Psychology*, *27*, 37–54.

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. *Advances in Experimental Social Psychology*, *24*, 201–234.

Ciesielski, B. G., Armstrong, T., Zald, D. H., & Olatunji, B. O. (2010). Emotion modulation of visual attention: Categorical and temporal characteristics. *PLOS ONE*, *5*(11), Article e13860. doi:10.1371/journal.pone.0013860

Cikara, M., Bruneau, E., Van Bavel, J. J., & Saxe, R. (2014). Their pain gives us pleasure: How intergroup dynamics shape empathic failures and counter-empathic responses. *Journal of Experimental Social Psychology*, *55*, 110–125.

Clore, G. L., & Huntsinger, J. R. (2007). How emotions inform judgment and regulate thought. *Trends in Cognitive Sciences*, *11*, 393–399. doi:10.1371/journal.pone.0013860

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo chamber or public sphere? Predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication*, *64*, 317–332.

Conger, K. (2019, October 30). Twitter will ban all political ads, C.E.O. Jack Dorsey says. *The New York Times*. Retrieved from https://www.nytimes.com/2019/10/30/technology/twitter-political-ads-ban.html

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 163–228). New York, NY: Oxford University Press.

Cosmides, L., Tooby, J., & Kurzban, R. (2003). Perceptions of race. *Trends in Cognitive Sciences*, *7*, 173–179.

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, *1*, 769–771.

Cushman, C. (2017, April 28). The Berkeley effect: Is this the end of civil discourse? *Fox News*. Retrieved from http://

www.foxnews.com/opinion/2017/04/28/berkeley-effect-christian-pastor-excluded-from-prayer-breakfast.html

de Hooge, I. E., Zeelenberg, M., & Breugelmans, S. M. (2007). Moral sentiments and cooperation: Differential influences of shame and guilt. *Cognition and Emotion*, *21*, 1025–1042.

DeSteno, D., Bartlett, M. Y., Baumann, J., Williams, L. A., & Dickens, L. (2010). Gratitude as moral sentiment: Emotion-guided cooperation in economic exchange. *Emotion*, *10*, 289–293.

Diener, E., Lusk, R., DeFour, D., & Flax, R. (1980). Deindividuation: Effects of group size, density, number of observers, and group member similarity on self-consciousness and disinhibited behavior. *Journal of Personality and Social Psychology*, *39*, 449–459.

Dijksterhuis, A., & Aarts, H. (2010). Goals, attention, and (un) consciousness. *Annual Review of Psychology*, *61*, 467–490.

Doosje, B., Branscombe, N. R., Spears, R., & Manstead, A. S. R. (1998). Guilty by association: When one's group has a negative history. *Journal of Personality and Social Psychology*, *75*, 872–886.

Duggan, M., & Smith, A. (2016, October 25). *The political environment on social media*. Retrieved from http://www.pewinternet.org/2016/10/25/political-content-on-social-media

Dunbar, R., & Dunbar, R. I. M. (1998). *Grooming, gossip, and the evolution of language*. Cambridge, MA: Harvard University Press.

Dunbar, R. I. M. (2004). Gossip in evolutionary perspective. *Review of General Psychology*, *8*, 100–110.

Dutta-Bergman, M. J. (2004). A descriptive narrative of healthy eating: A social marketing approach using psychographics in conjunction with interpersonal, community, mass media and new media activities. *Health Marketing Quarterly*, *20*, 81–101.

Egyed, K., Király, I., & Gergely, G. (2013). Communicating shared knowledge in infancy. *Psychological Science*, *24*, 1348–1353.

Ekman, P. (1994). Antecdent events and emotion metaphors. In P. Ekman & R. Davidson (Eds.), *The nature of emotion: Fundamental questions* (pp. 146–149). New York, NY: Oxford University Press.

Ellemers, N., Spears, R., & Doosje, B. (2002). Self and social identity. *Annual Review of Psychology*, *53*, 161–186.

Ellsworth, P., & Scherer, K. (2003). Appraisal processes in emotion. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 572–595). New York, NY: Oxford University Press.

Enli, G. (2017). Twitter as arena for the authentic outsider: Exploring the social media campaigns of Trump and Clinton in the 2016 US presidential election. *European Journal of Communication*, *32*, 50–61.

Errasti, J., Amigo, I., & Villadangos, M. (2017). Emotional uses of Facebook and Twitter: Its relation with empathy, narcissism, and self-esteem in adolescence. *Psychological Reports*, *120*, 997–1018.

Evans, S. K., Pearce, K. E., Vitak, J., & Treem, J. W. (2017). Explicating affordances: A conceptual framework for understanding affordances in communication research. *Journal of Computer-Mediated Communication*, *22*, 35–52.

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.

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

Fan, R., Zhao, J., Chen, Y., & Xu, K. (2014). Anger is more influential than joy: Sentiment correlation in Weibo. *PLOS ONE*, *9*(10), Article e110184. doi:10.1371/journal.pone.0110184

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*, 63–87.

Festinger, L., Pepitone, A., & Newcomb, T. (1952). Some consequences of de-individuation in a group. *Journal of Abnormal and Social Psychology*, *47*, 382–389.

Firestone, C., & Scholl, B. J. (2015). Enhanced visual awareness for morality and pajamas? Perception vs. memory in "top-down" effects. *Cognition*, *136*, 409–416.

Fischer, A., & Manstead, A. (2004). Social influences on the emotion process. *European Review of Social Psychology*, *14*, 171–201.

Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology*, *38*, 889–906.

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

Fitz, N., Kushlev, K., Jagannathan, R., Lewis, T., Paliwal, D., & Ariely, D. (2019). Batching smartphone notifications can improve well-being. *Computers in Human Behavior*, *101*, 84–94.

Fordy, T. (2014, July 2). Is there a misogynist inside every man? *The Telegraph*. Retrieved from https://www.telegraph.co.uk/men/thinking-man/10924854/Is-there-a-misogynist-inside-every-man.html

Frijda, N. (1986). *The emotions: Studies in emotion and social interaction*. Cambridge, England: Cambridge University Press.

Frimer, J. A., Boghrati, R., Haidt, J., Graham, J., & Dehgani, M. (2019). *Moral foundations dictionary for linguistic analyses 2.0*. Charlottesville, VA: Center for Open Science. doi:10.17605/OSF.IO/EZN37

Gantman, A. P., & Van Bavel, J. J. (2014). The moral pop-out effect: Enhanced perceptual awareness of morally relevant stimuli. *Cognition*, *132*, 22–29.

Gantman, A. P., & Van Bavel, J. J. (2015). Moral perception. *Trends in Cognitive Sciences*, *19*, 631–633.

Gantman, A. P., & Van Bavel, J. J. (2016). Exposure to justice diminishes moral perception. *Journal of Experimental Psychology: General*, *145*, 1728–1739.

Goldenberg, A., Halperin, E., van Zomeren, M., & Gross, J. J. (2016). The process model of group-based emotion: Integrating intergroup emotion and emotion regulation perspectives. *Personality and Social Psychology Review*, *20*, 118–141.

Goodwin, G. P., & Darley, J. M. (2012). Why are some moral beliefs perceived to be more objective than others? *Journal of Experimental Social Psychology*, *48*, 250–256.

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*, 1029–1046.

Granot, Y., Balcetis, E., Schneider, K. E., & Tyler, T. R. (2014). Justice is not blind: Visual attention exaggerates effects of group identification on legal punishment. *Journal of Experimental Psychology: General*, *143*, 2196–2208.

Greene, J. D. (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *The neuroscience of morality: Emotion, brain disorders, and development* (pp. 35–80). Cambridge, MA: MIT Press.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*, 389–400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*, 2105–2108.

Grönlund, K., Herne, K., & Setälä, M. (2015). Does enclave deliberation polarize opinions? *Political Behavior*, *37*, 995–1020.

Gross, J. J. (1999). Emotion and emotion regulation. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 525–552). New York, NY: Guilford Press.

Grubbs, J. B., Warmke, B., Tosi, J., & James, A. S. (2019). Grandstanding moral grandstanding in public discourse: Status-seeking motives as a potential explanatory mechanism in predicting conflict. *PLOS ONE*, *14*(10), Article 223749. doi:10.1371/journal.pone.0223749

Grusec, J. E., & Goodnow, J. J. (1994). Impact of parental discipline methods on the child's internalization of values: A reconceptualization of current points of view. *Developmental Psychology*, *30*, 4–19.

Guerini, M., & Staiano, J. (2015). Deep feelings: A massive cross-lingual study on the relation between emotions and virality. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 299–305). doi:10.1145/2740908.2743058

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, *5*(1), Article eaau4586. doi:10.1126/sciadv.aau4586

Hafer, C. L. (2000). Do innocent victims threaten the belief in a just world? Evidence from a modified Stroop task. *Journal of Personality and Social Psychology*, *79*, 165–173.

Haidt, J. (2000). The positive emotion of elevation. *Prevention & Treatment*, *3*(1), Article 3c. doi:10.1037/1522-3736.3.1.33c

Haidt, J. (2003). The moral emotions. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 572–595). New York, NY: Oxford University Press.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. New York, NY: Pantheon Books.

Haidt, J., Rozin, P., McCauley, C., & Imada, S. (1997). Body, psyche, and culture: The relationship between disgust and morality. *Psychology and Developing Societies*, *9*, 107–131.

Hansen, L.K., Arvidsson, A., Nielsen, F.A., Colleoni, E., & Etter, M. (2011) Good friends, bad news—affect and virality in Twitter. In J. J. Park, L. T. Yang, & C. Lee (Eds.), *Future information technology* (pp. 34–43). Amsterdam, The Netherlands: Springer. doi:10.1007/978-3-642-22309-9_5

Haslam, S. A., Reicher, S. D., & Platow, M. J. (2010). *The new psychology of leadership: Identity, influence and power*. New York, NY: Psychology Press.

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1993). Emotional contagion. *Current Directions in Psychological Science*, *2*, 96–100.

Heath, C., Bell, C., & Sternberg, E. (2001). Emotional selection in memes: The case of urban legends. *Journal of Personality and Social Psychology*, *81*, 1028–1041.

Heimbach, I., Schiller, B., Strufe, T., & Hinz, O. (2015). Content virality on online social networks: Empirical evidence from Twitter, Facebook, and Google+ on German news websites. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media* (pp. 39–47). New York, NY: Association for Computing Machinery. doi:10.1145/2700171.2791032

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*, 40–60.

Ho, M. K., MacGlashan, J., Littman, M. L., & Cushman, F. (2017). Social is special: A normative framework for teaching with and learning from evaluative feedback. *Cognition*, *167*, 91–106.

Hochreiter, R., & Waldhauser, C. (2014). *The role of emotions in propagating brands in social networks*. Retrieved from https://arxiv.org/abs/1409.4617

Hogg, M. A. (2007). Uncertainty-identity theory. *Advances in Experimental Social Psychology*, *39*, 69–126.

Hornsey, M. J. (2008). Social identity theory and self-categorization theory: A historical review. *Social and Personality Psychology Compass*, *2*, 204–222.

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

Iyer, A., Schmader, T., & Lickel, B. (2007). Why individuals protest the perceived transgressions of their country: The role of anger, shame, and guilt. *Personality and Social Psychology Bulletin*, *33*, 572–587.

Izard, C. E. (1977). *Human emotions*. New York, NY: Springer Science & Business Media.

Izuma, K., Saito, D. N., & Sadato, N. (2008). Processing of social and monetary rewards in the human striatum. *Neuron*, *58*, 284–294.

Jacquet, J. (2016). *Is shame necessary?: New uses for an old tool*. New York, NY: Vintage.

Jaidka, K., Zhou, A., & Lelkes, Y. (2019). Brevity is the soul of Twitter: The constraint affordance and political discussion. *Journal of Communication*, *69*, 345–372.

Jetten, J., Branscombe, N. R., Schmitt, M. T., & Spears, R. (2001). Rebels with a cause: Group identification as a response to perceived discrimination from the mainstream. *Personality and Social Psychology Bulletin*, *27*, 1204–1213.

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

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, *530*, 473–476.

Jordan, J. J., & Rand, D. G. (2020). Signaling when no one is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interactions. *Journal of Personality and Social Psychology*, *118*, 57–88.

Kay, A. C., Gaucher, D., McGregor, I., & Nash, K. (2010). Religious belief as compensatory control. *Personality and Social Psychology Review*, *14*, 37–48.

Kay, A. C., & Jost, J. T. (2003). Complementary justice: Effects of "poor but happy" and "poor but honest" stereotype exemplars on system justification and implicit activation of the justice motive. *Journal of Personality and Social Psychology*, *85*, 823–837.

Keil, A., & Ihssen, N. (2004). Identification facilitation for emotionally arousing verbs during the attentional blink. *Emotion*, *4*, 23–35.

Keltner, D., & Haidt, J. (1999). Social functions of emotions at four levels of analysis. *Cognition and Emotion*, *13*, 505–521.

Kissler, J., Herbert, C., Peyk, P., & Junghofer, M. (2007). Buzzwords: Early cortical responses to emotional words during reading: Research report. *Psychological Science*, *18*, 475–480.

Kissler, J., Herbert, C., Winkler, I., & Junghofer, M. (2009). Emotion and attention in visual word processing: An ERP study. *Biological Psychology*, *80*, 75–83.

Klein, R. M. (2000). Inhibition of return. *Trends in Cognitive Sciences*, *4*, 138–147.

Kohlberg, L., Levine, C., & Hewer, A. (1983). Moral stages: A current formulation and a response to critics. *Contributions to Human Development*, *10*, 174.

Kollanyi, B., Howard, P. N., & Woolley, S. C. (2016). Bots and automation over Twitter during the first U.S. election. *COMPROP Data Memo*, *1*, 1–4.

Kramer, A., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences, USA*, *111*, 8788–8790.

Kugler, M., Jost, J. T., & Noorbaloochi, S. (2014). Another look at moral foundations theory: Do authoritarianism and social dominance orientation explain liberal-conservative differences in "moral" intuitions? *Social Justice Research*, *27*, 413–431.

Lazarus, R. S. (1966). *Psychological stress and the coping process*. New York, NY: McGraw-Hill.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., . . . Zittrain, J. L. (2018). The science of fake news. *Science*, *359*, 1094–1096.

Ledgerwood, A., & Callahan, S. P. (2012). The social side of abstraction: Psychological distance enhances conformity to group norms. *Psychological Science*, *23*, 907–913.

Lerner, M. J., & Miller, D. T. (1978). Just world research and the attribution process: Looking back and ahead. *Psychological Bulletin*, *85*, 1030–1051.

Levendusky, M. S., & Malhotra, N. (2016). (Mis)perceptions of partisan polarization in the American public. *Public Opinion Quarterly*, *80*, 378–391.

Lindquist, K. A., MacCormack, J. K., & Shablack, H. (2015). The role of language in emotion: Predictions from psychological constructionism. *Frontiers in Psychology*, *6*, Article 444. doi:10.3389/fpsyg.2015.00444

Livingstone, A. G., Shepherd, L., Spears, R., & Manstead, A. S. R. (2016). "Fury, us": Anger as a basis for new group self-categories. *Cognition and Emotion*, *30*, 183–192.

Locke, K. D., & Nekich, J. C. (2000). Agency and communion in naturalistic social comparison. *Personality and Social Psychology Bulletin*, *26*, 864–874.

Lotan, G., Graeff, E., Ananny, M., Gaffney, D., Pearce, I., & Boyd, D. (2011). The Arab Spring—the revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian Revolutions. *International Journal of Communication*, *5*, 1375–1405.

Loureiro, M. L., & Lotade, J. (2005). Do fair trade and eco-labels in coffee wake up the consumer conscience? *Ecological Economics*, *53*, 129–138.

Lu, Z., Wen, Y., & Cao, G. (2014). Information diffusion in mobile social networks: The speed perspective. In *Proceedings—IEEE INFOCOM* (pp. 1932–1940). New York, NY: Institute of Electrical and Electronics Engineers. doi:10.1109/INFOCOM.2014.6848133

Luttrell, A., Petty, R. E., Briñol, P., & Wagner, B. C. (2016). Making it moral: Merely labeling an attitude as moral increases its strength. *Journal of Experimental Social Psychology*, *65*, 82–93.

MacCarthy, R. (2016, June 23). The average Twitter user now has 707 followers [Blog post]. Retrieved from https://kickfactory.com/blog/average-twitter-followers-updated-2016

Mackie, D. M., Devos, T., & Smith, E. R. (2000). Intergroup emotions: Explaining offensive action tendencies in an intergroup context. *Journal of Personality and Social Psychology*, *79*, 602–616.

Mackie, D. M., Silver, L. A., & Smith, E. R. (2004). Intergroup emotions: Emotion as an intergroup phenomenon. In L. Z. Tiedens & C. W. Leach (Eds.), *The social life of emotions* (pp. 227–245). New York, NY: Cambridge University Press.

Matheson, K., & Zanna, M. P. (1988). The impact of computer-mediated communication on self-awareness. *Computers in Human Behavior*, *4*, 221–233.

Maza, C. (2018, September 21). Why every social media site is a dumpster fire. *Vox*. Retrieved from https://www.vox.com/2018/9/21/17886400/strikethrough-social-media-dumpster-fire-trolls-tribalism

McCullough, M. E., Emmons, R. A., Kilpatrick, S. D., & Larson, D. B. (2001). Is gratitude a moral affect? *Psychological Bulletin*, *127*, 249–266.

McGarty, C., Thomas, E. F., Lala, G., Smith, L. G. E., & Bliuc, A. M. (2014). New technologies, new identities, and the growth of mass opposition in the Arab Spring. *Political Psychology*, *35*, 725–740.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*, 415–444.

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*, Article 439. doi:10.3389/fnhum.2013.00439

Milders, M., Sahraie, A., Logan, S., & Donnellon, N. (2006). Awareness of faces is modulated by their emotional meaning. *Emotion*, *6*, 10–17.

Miller, W. I. (1997). *The anatomy of disgust*. Cambridge, MA: Harvard University Press.

Mooijman, M., Hoover, J., Lin, Y., Ji, H., & Dehghani, M. (2018). Moralization in social networks and the emergence of violence during protests. *Nature Human Behaviour*, *2*, 389–396.

Most, S. B., Smith, S. D., Cooter, A. B., Levy, B. N., & Zald, D. H. (2007). The naked truth: Positive, arousing distractors impair rapid target perception. *Cognition and Emotion*, *21*, 964–981.

Most, S. B., & Wang, L. (2011). Dissociating spatial attention and awareness in emotion-induced blindness. *Psychological Science*, *22*, 300–305.

Mrowicki, M. (2015). *Decline in civil discourse threatens the heart of our democracy*. Retrieved from http://www.commonsnews.org/site/sitenext/story.php?articleno=12143&page=1

Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, *39*, 629–649.

Munger, K., & Phillips, J. (2019). *A supply and demand framework for YouTube politics*. Retrieved from https://osf.io/73jys

Mutz, D. C. (2002). Cross-cutting social networks: Testing democratic theory in practice. *American Political Science Review*, *96*, 111–126.

Niedenthal, P. M., Tangney, J. P., & Gavanski, I. (1994). "If only I weren't" versus "if only I hadn't": Distinguishing shame and guilt in counterfactual thinking. *Journal of Personality and Social Psychology*, *67*, 585–595.

Nitschke, J. B., Nelson, E. E., Davidson, R. J., Oakes, T. R., Rusch, B. D., & Fox, A. S. (2004). Orbitofrontal cortex tracks positive mood in mothers viewing pictures of their newborn infants. *NeuroImage*, *21*, 583–592.

Noel, J. (2017, January 27). Facts are not dead, but civil discourse is on life support. *Chicago Tribune*. Retrieved from http://www.chicagotribune.com/news/opinion/commentary/ct-trump-truth-facts-20170127-story.html

Nussbaum, M. C. (2016). *Anger and forgiveness: Resentment, generosity, justice*. New York, NY: Oxford University Press.

Ohman, A., Flykt, A., & Esteves, F. (2001). Emotion drives attention: Detecting the snake in the grass. *Journal of Experimental Psychology: General*, *130*, 466–478.

Öhman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review*, *108*, 483–522.

Ortigue, S., Michel, C. M., Murray, M. M., Mohr, C., Carbonnel, S., & Landis, T. (2004). Electrical neuroimaging reveals early generator modulation to emotional words. *NeuroImage*, *21*, 1242–1251.

Parkinson, B. (2011). Interpersonal emotion transfer: Contagion and social appraisal. *Social and Personality Psychology Compass*, *5*, 428–439.

Pärnamets, P., Johansson, P., Hall, L., Balkenius, C., Spivey, M. J., & Richardson, D. C. (2015). Biasing moral decisions by exploiting the dynamics of eye gaze. *Proceedings of the National Academy of Sciences, USA*, *112*, 4170–4175.

Pärnamets, P., Reinero, D. A., Pereira, A., & Van Bavel, J. J. (2019). Identity leadership: Managing perceptions of conflict for collective action. *Behavioral and Brain Sciences*, *42*, Article e136. doi:10.1017/S0140525X19000876

Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences, USA*, *116*, 2521–2526.

Penrose, N. (2018 September 27). Women are thanking Dr. Ford for her bravery on Twitter. *Elle*. Retrieved from https://www.elle.com/culture/career-politics/a23493387/twitter-reactions-christine-ford-kavanaugh-hearing

Peters, K., & Kashima, Y. (2007). From social talk to social action: Shaping the social triad with emotion sharing. *Journal of Personality and Social Psychology*, *93*, 780–797.

Peters, K., & Kashima, Y. (2015). A multimodal theory of affect diffusion. *Psychological Bulletin*, *141*, 966–992.

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*, 117–128.

Popken, B. (2018, April 19). As algorithms take over, YouTube's recommendations highlight a human problem. *NBC News*. Retrieved from https://www.nbcnews.com/tech/social-media/algorithms-take-over-youtube-s-recommendations-highlight-human-problem-n867596

Postmes, T., & Spears, R. (1998). Deindividuation and antinormative behavior: A meta-analysis. *Psychological Bulletin*, *123*, 238–259.

Postmes, T., Spears, R., & Lea, M. (1998). Breaching or building social boundaries? SIDE-effects of computer-mediated communication. *Communication Research*, *25*, 689–715.

Postmes, T., Spears, R., & Lea, M. (2000). The formation of group norms in computer-mediated communication. *Human Communication Research*, *26*, 341–371.

Pratto, F., & John, O. P. (1991). Automatic vigilance: The attention-grabbing power of negative social information. *Journal of Personality and Social Psychology*, *61*, 380–391.

Prentice, D. A., & Miller, D. T. (1993). Pluralistic ignorance and alcohol use on campus: Some consequences of misperceiving the social norm. *Journal of Personality and Social Psychology*, *64*, 243–256.

Quercia, D., Ellis, J., Capra, L., & Crowcroft, J. (2011). In the mood for being influential on Twitter. In *IEEE Third International Conference on Privacy, Security, Risk and*

Trust and 2011 IEEE Third International Conference on Social Computing (pp. 307–314). New York, NY: Institute of Electrical and Electronics Engineers. doi:10.1109/PASSAT/SocialCom.2011.27

Reicher, S. (1982). The determination of collective behaviour. In H. Tajfel (Eds.), *Social identity and intergroup relations* (pp. 41–83). Cambridge, England: Cambridge University Press.

Reicher, S. D. (1984). Social influence in the crowd: Attitudinal and behavioural effects of de-individuation in conditions of high and low group salience. *British Journal of Social Psychology*, *23*, 341–350.

Rime, B., Mesquita, B., Philippot, P., & Boca, S. (1991). Beyond the emotional event: Six studies on the social sharing of emotion. *Cognition and Emotion*, *5*, 435–465.

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

Ronson, J. (2016). *So you've been publicly shamed*. New York, NY: Riverhead Books.

Rose-Stockwell, T. (2017, July 14). This is how your fear and outrage are being sold for profit. *Medium*. Retrieved from https://medium.com/the-mission/the-enemy-in-our-feeds-e86511488de

Rose-Stockwell, T. (2018 April 13). How to design better social media. *Medium*. Retrieved from https://medium.com/s/story/how-to-fix-what-social-media-has-broken-cb0b2737128

Rost, K., Stahel, L., & Frey, B. S. (2016). Digital social norm enforcement: Online firestorms in social media. *PLOS ONE*, *11*(6), Article 155923. doi:10.1371/journal.pone.0155923

Rozin, P. (1999). The process of moralization. *Psychological Science*, *10*, 218–221.

Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). The CAD triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology*, *76*, 574–586.

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience*, *15*, 549–562.

Sage, K. D., & Baldwin, D. (2011). Disentangling the social and the pedagogical in infants' learning about tool-use. *Social Development*, *20*, 825–844.

Salerno, J. M., & Peter-Hagene, L. C. (2013). The interactive effect of anger and disgust on moral outrage and judgments. *Psychological Science*, *24*, 2069–2078.

Sawaoka, T., & Monin, B. (2018). The paradox of viral outrage. *Psychological Science*, *29*, 1665–1678.

Schmidt, A. L., Zollo, F., Scala, A., Betsch, C., & Quattrociocchi, W. (2018). Polarization of the vaccination debate on Facebook. *Vaccine*, *36*, 3606–3612.

Schmitt, M., Baumert, A., Gollwitzer, M., & Maes, J. (2010). The Justice Sensitivity Inventory: Factorial validity, location in the personality facet space, demographic pattern, and normative data. *Social Justice Research*, *23*, 211–238.

Sheikh, S., & Janoff-Bulman, R. (2010). The "shoulds" and "should nots" of moral emotions: A self-regulatory perspective on shame and guilt. *Personality and Social Psychology Bulletin*, *36*, 213–224.

Sherif, M. (1961). *Intergroup conflict and cooperation: The Robbers Cave experiment*. Norman, OK: University Book Exchange.

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*, 1027–1035.

Shirazi, A. S., Henze, N., Dingler, T., Pielot, M., Weber, D., & Schmidt, A. (2014). Large-scale assessment of mobile notifications. In *CHI '14: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3055–3064). New York, NY: Association for Computing Machinery. doi:10.1145/2556288.2557189

Siegel, A., & Badaan, V. (2020). *#No2Sectarianism: Experimental approaches to reducing sectarian hate speech online*. Manuscript submitted for publication.

Siegel, J. Z., Mathys, C., Rutledge, R. B., & Crockett, M. J. (2018). Beliefs about bad people are volatile. *Nature Human Behaviour*, *2*, 750–756.

Sillito, D. (2016 November 14). Donald Trump: How the media created the president. *BBC*. Retrieved from https://www.bbc.com/news/entertainment-arts-37952249

Simon, B., & Klandermans, B. (2001). Politicized collective identity. *American Psychologist*, *56*, 319–331.

Simon, H. A. (1996). Designing organizations for an information-rich world. *International Library of Critical Writings in Economics*, *70*, 187–202

Singer, M. (1961). *Generalization in ethics*. New York, NY: Knopf.

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*, 895–917.

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin*, *105*, 131–142.

Skrandies, W. (1998). Evoked potential correlates of semantic meaning—A brain mapping study. *Cognitive Brain Research*, *6*, 173–183.

Smith, A. (2014, February 3). *What people like and dislike about Facebook*. Retrieved from https://www.pewresearch.org/fact-tank/2014/02/03/what-people-like-dislike-about-facebook

Smith, E. R., Seger, C. R., & Mackie, D. M. (2007). Can emotions be truly group level? Evidence regarding four conceptual criteria. *Journal of Personality and Social Psychology*, *93*, 431–446.

Spring, V., Cameron, D., & Cikara, M. (2018). The upside of outrage. *Trends in Cognitive Sciences*, *22*, 1067–1069.

Srinivasan, A. (2018). The aptness of anger. *Journal of Political Philosophy*, *26*, 123–144.

Statistica. (2018). *Number of monthly active Twitter users worldwide from 1st quarter 2010 to 2nd quarter 2018 (in millions)*. Retrieved from https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users

Sterling, J., & Jost, J. T. (2018). Moral discourse in the Twitterverse: Effects of ideology and political sophistication on language use among U.S. citizens and members of Congress. *Journal of Language and Politics*, *17*, 195–221.

Stieglitz, S., & Dang-Xuan, L. (2012). Political communication and influence through microblogging—An empirical analysis of sentiment in Twitter messages and retweet behavior. In *2012 45th Hawaii International Conference on System Sciences* (pp. 3500–3509). New York, NY: Institute of Electrical and Electronics Engineers. doi:10.1109/HICSS.2012.476

Stieglitz, S., & Dang-Xuan, L. (2013). Emotions and information diffusion in social media—Sentiment of microblogs and sharing behavior. *Journal of Management Information Systems*, *29*, 217–248.

Strohminger, N., & Nichols, S. (2015). Neurodegeneration and identity. *Psychological Science*, *26*, 1469–1479.

Suter, R. S., & Hertwig, R. (2011). Time and moral judgment. *Cognition*, *119*, 454–458.

Sutton, K. (2016, January 8). Upworthy changes course; lays off 14 to focus resources on video. *Politico*. Retrieved from https://www.politico.com/media/story/2016/01/upworthy-changes-course-lays-off-14-to-focus-resources-on-video-004345

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33–47). Monterey, CA: Wadsworth.

Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behavior. *Psychology of Intergroup Relations*, *5*, 7–24.

Tangney, J. P., Miller, R. S., Flicker, L., & Barlow, D. H. (1996). Are shame, guilt, and embarrassment distinct emotions? *Journal of Personality and Social Psychology*, *70*, 1256–1269.

Tangney, J. P., & Tracey, J. L. (2012). Self-conscious emotions. In M. R. Leary & J. P. Tangney (Eds.), *Handbook of self and identity* (2nd ed., pp. 446–478). New York, NY: Guilford Press.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*, 24–54.

Tetlock, P. E., Kristel, O. V., Elson, S. B., Green, M. C., & Lerner, J. S. (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology*, *78*, 853–870.

Timberg, C., Dwoskin, E., Entous, A., & Demirjian, K. (2017, November 1). Russian ads, now publicly released, show sophistication of influence campaign. *The Washington Post*. Retrieved from https://www.washingtonpost.com/business/technology/russian-ads-now-publicly-released-show-sophistication-of-influence-campaign/2017/11/01/d26aead2-bf1b-11e7-8444-a0d4f04b89eb_story.html

Tracy, J. L., & Robins, R. W. (2004). Putting the self into self-conscious emotions: A theoretical model. *Psychological Inquiry*, *5*, 103–125

Tracy, J. L., & Robins, R. W. (2006). Appraisal antecedents of shame and guilt: Support for a theoretical model. *Journal of Personality and Social Psychology*, *32*, 1339–1351

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review*, *117*, 440–463.

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory*. New York, NY: Blackwell.

Turner, J. C., Oakes, P. J., Haslam, S. A., & McGarty, C. (2007). Self and collective: Cognition and social context. *Personality and Social Psychology Bulletin*, *20*, 454–463.

Uhlmann, E. L., Zhu, L. L., & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition*, *126*, 326–334.

Vaglanos, A. (2017, October 18). In response to #MeToo, men are tweeting #HowIWillChange. *HuffPost*. Retrieved from https://www.huffingtonpost.com/entry/in-response-to-metoo-men-are-tweeting-howiwillchange_us_59e79bd3e4b00905bdae455d

Valenzuela, S., Piña, M., & Ramírez, J. (2017). Behavioral effects of framing on social media users: How conflict, economic, human interest, and morality frames drive news sharing. *Journal of Communication*, *67*, 803–826.

Van Bavel, J. J., & Cunningham, W. A. (2012). A social identity approach to person memory: Group membership, collective identification, and social role shape attention and memory. *Personality and Social Psychology Bulletin*, *38*, 1566–1578.

Van Bavel, J. J., Packer, D. J., Haas, I. J., & Cunningham, W. A. (2012). The importance of moral construal: Moral versus non-moral construal elicits faster, more extreme, universal evaluations of the same actions. *PLOS ONE*, *7*(11), Article e48693. doi:10.1371/journal.pone.0048693

Van Bavel, J. J., & Pereira, A. (2018). The partisan brain: An identity-based model of political belief the role of identity in political belief. *Trends in Cognitive Sciences*, *22*, 213–224.

Van Der Linden, S. (2017). The nature of viral altruism and how to make it stick. *Nature Human Behaviour*, *1*(3), 1–4.

Van Kleef, G. A. (2009). How emotions regulate social life: The emotions as social information (EASI) model. *Current Directions in Psychological Science*, *18*, 184–188.

Van Kleef, G. A. (2017). The social effects of emotions are functionally equivalent across expressive modalities. *Psychological Inquiry*, *28*, 211–216.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*, 1146–1151.

Vredenburgh, C., Kushnir, T., & Casasola, M. (2015). Pedagogical cues encourage toddlers' transmission of recently demonstrated functions to unfamiliar adults. *Developmental Science*, *18*, 645–654.

Wade, J. (2017, September 13). *Social Media Week London 2017 recap*. Retrieved from https://www.smartinsights.com/social-media-marketing/social-media-week-london-2017-recap

Williams, J. (2018). *Stand out of our light: Freedom and resistance in the attentional economy*. New York, NY: Cambridge University Press.

Wohl, M. J. A., Branscombe, N. R., & Klar, Y. (2006). Collective guilt: Emotional reactions when one's group has done

wrong or been wronged. *European Review of Social Psychology*, *17*, 1–37.

Xiao, Y. J., Coppin, G., & Van Bavel, J. J. (2016). Perceiving the world through group-colored glasses: A perceptual model of intergroup relations. *Psychological Inquiry*, *27*, 255–274.

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

Yudkin, D. A., Rothmund, T., Twardawski, M., Thalla, N., & Van Bavel, J. J. (2016). Reflexive intergroup bias in third-party punishment. *Journal of Experimental Psychology: General*, *145*, 1448–1459.

Zaki, J. (2019 August 6). The technology of kindness. *Scientific American*. Retrieved from https://www.scientificamerican.com/article/the-technology-of-kindness

Zaki, J., & Williams, W. (2013). Interpersonal emotion regulation. *Emotion*, *13*, 803–810.