ANNUAL REVIEWS

Annual Review of Psychology

Social Media and Morality

Jay J. Van Bavel,[1,2] Claire E. Robertson,[1]
Kareena del Rosario,[1] Jesper Rasmussen,[3]
and Steve Rathje[1]

[1] Department of Psychology, New York University, New York, NY, USA;
email: jay.vanbavel@nyu.edu, cer493@nyu.edu, kdr321@nyu.edu, srathje@alumni.stanford.edu

[2] Department of Strategy and Management, Norwegian School of Economics, Bergen, Norway

[3] Department of Political Science, Aarhus University, Aarhus, Denmark; email: jr@ps.au.dk

Annu. Rev. Psychol. 2024. 75:22.1–22.30

The *Annual Review of Psychology* is online at psych.annualreviews.org

https://doi.org/10.1146/annurev-psych-022123-110258

Keywords

morality, social media, emotions, outrage, identity, politics

Abstract

Nearly five billion people around the world now use social media, and this number continues to grow. One of the primary goals of social media platforms is to capture and monetize human attention. One means by which individuals and groups can capture attention and drive engagement on these platforms is by sharing morally and emotionally evocative content. We review a growing body of research on the interrelationship of social media and morality as well its consequences for individuals and society. Moral content often goes viral on social media, and social media makes moral behavior (such as punishment) less costly. Thus, social media often acts as an accelerant for existing moral dynamics, amplifying outrage, status seeking, and intergroup conflict while also potentially amplifying more constructive facets of morality, such as social support, prosociality, and collective action. We discuss trends, heated debates, and future directions in this emerging literature.

Contents

INTRODUCTION

> The real problem of humanity is the following: We have Paleolithic emotions, medieval institutions and godlike technology. And it is terrifically dangerous, and it is now approaching a point of crisis overall.
>
> —Edward O. Wilson (cited in Harv. Mag. 2009)

The growth of social media use has been a prominent trend over the past decade, with platforms such as Facebook, Twitter, Instagram, and TikTok attracting billions of active users globally. As of 2023, there are nearly 5 billion active social media users around the globe and the average user spends 2 hours and 27 minutes on social media per day (Statista 2023). According to Facebook, the average person scrolls through roughly 300 feet of content every day on their mobile device—roughly the height of the Statue of Liberty (Morant 2018). The massive growth in social media has raised serious questions about how human psychology shapes our online behavior as well as how the online environment shapes our psychology. This article reviews the literature on the interrelationship between social media and morality.

Social media refers to Internet-based platforms for mass personal communication that facilitate interactions among users and derive their value primarily from user-generated content (Carr & Hayes 2015). One of the primary goals of social media platforms (and of their leaders, employers, and shareholders) is to capture and monetize human attention (usually by selling advertisements). Yet, the limits of human attention create an attention economy, whereby individuals must prioritize which information they will attend to while platforms, advertisers, and other users compete for that scarce attention. This leads platforms to employ personalized algorithms, compelling design features, and prominent influencers to capture and retain users' attention. Competing in this attention economy can lead people to create and express exaggerated beliefs and carefully curated content designed to capture attention rather than reflect reality or benefit humanity (Fisher 2022b). The pull of the attention economy is thus central to understanding the psychology and behavior of social media users.

One means by which individuals can capture attention and drive engagement is by sharing morally and emotionally evocative content (see Brady et al. 2020a,b). Our attention to moral issues is the result of more than 6 million years of evolution in which survival often hinged on the ability of individuals and coalitions to effectively navigate their social groups (Boehm 2012, Krebs 2008). Even today, transgressing moral norms can result in public shaming, punishment, and professional censure, while affirming community values or moral heroism can immediately bolster one's social status (Haidt 2007, Salerno & Peter-Hagene 2013). Moralistic punishment can be a force for good, increasing cooperation by holding bad actors accountable; but punishment can also increase social conflict by dehumanizing others and even escalate into violent conflict (Fincher & Tetlock 2016). As such, morality plays a central role in human sociality, from regulating social behavior to mobilizing collective action.

Social media appears to accelerate and inflame these facets of moral psychology, fostering and rewarding norms of outrage (Brady et al. 2021). Indeed, people now encounter more acts related to morality and experience stronger feelings of moral outrage from online content than from print media, TV, and radio combined (Crockett 2017). Our online exposure to (im)morality might reflect the fact that moral content captures attention (Brady et al. 2020b), and social media makes moral behavior (such as punishment) less costly. This has raised the question, If moral outrage is a fire, is the Internet like gasoline? (Crockett 2017, p. 771). Our review of the literature suggests that social media often acts like an accelerant for existing moral dynamics—amplifying negative facets of morality such as outrage, harassment, status seeking, and intergroup conflict as well as some positive aspects of morality, such as social support, prosociality, and collective action.

Whether social media has had a net positive or net negative effect on the well-being of individuals, groups, and society as a whole is hotly contested (Lorenz-Spreen et al. 2023; Orben & Przybylski 2019; Orben et al. 2022; J. Haidt & C. Bail, unpublished manuscript). On one hand, social media has allowed people to build personal and professional relationships (Rajkumar et al. 2022), increase their knowledge (Allcott et al. 2020), and mobilize for social change (Zhuravskaya et al. 2020). On the other hand, prolonged exposure to the constant barrage of information on social media may lead to reduced well-being, greater stress, and social conflict (Allcott et al. 2020, Kross et al. 2021). However, the impact on well-being is still a source of academic debate. The algorithms and design features used by social media platforms may also accelerate exposure to misinformation (Vosoughi et al. 2018), facilitate harassment (Marwick 2021), reinforce echo chambers (Brady et al. 2017, Cinelli et al. 2021), and lead to political polarization (Allcott et al. 2020, Van Bavel et al. 2021b). Moreover, recent events suggest that moralized rhetoric on social media may inflame real world violence (Mooijman et al. 2018) as well as pose serious threats to public health and the functioning of democracy (Simchon et al. 2022).

Thus, the interaction between our ancient moral instincts and modern social media platforms may provide a unique set of opportunities and challenges for individuals and society (Bak-Coleman et al. 2021). In this review, we describe the psychological function and appeal of morality and how these aspects are harnessed and monetized by individuals, political actors, and social media companies. We then review the impact of our moral psychology on a broad array of individual and societal issues related to social media. Finally, we discuss the current and future directions of research on social media and morality (such as the need for more research across platforms and cultural contexts). This article aims to provide a state-of-the-art overview of how the interplay between morality and social media influences individuals, groups, and societies.

THE ORIGINS OF HUMAN MORALITY

At its core, moral psychology refers to people's beliefs about right and wrong (Ellemers & van den Bos 2012). Beliefs or issues become moralized when they are construed in terms of the interests or good of a unit larger than the individual (e.g., society, culture, one's social network; Haidt 2003, Rozin 1999). Thus, how people construe beliefs or issues is crucial to moralization and changes how people evaluate actions (Jarudi et al. 2008, Rozin 1999, Van Bavel et al. 2012). Moral beliefs are heavily influenced by social identities, such as political and religious identities (Cohen 2015, Haidt & Graham 2007, Tetlock et al. 2000). Moreover, morality informs our political opinions, and our political beliefs are often moralized (Skitka & Morgan 2014, Skitka & Washburn 2016).

The role of identity and morality appears to be particularly potent on social media, where people can easily signal their identity or beliefs to their social network (Brady et al. 2020a, Van Bavel & Packer 2021). For instance, messages with moral-emotional words (e.g., "hate") are far more likely to be shared on social media than messages about similar political topics that do not include these terms (Brady et al. 2017). Moreover, this language is associated with polarized sharing, such that partisans are more likely to share messages from fellow in-group members when they use moral emotional language (see **Figure 1**). Therefore, this review discusses the relationship between morality and politics, as politics is closely related to morality and identity (Simon & Klandermans 2001).

Beliefs that are based on moral conviction, or rooted in moral values, are stronger and more resistant to change than nonmoral beliefs (Aramovich et al. 2012, Hornsey et al. 2003, Skitka et al.

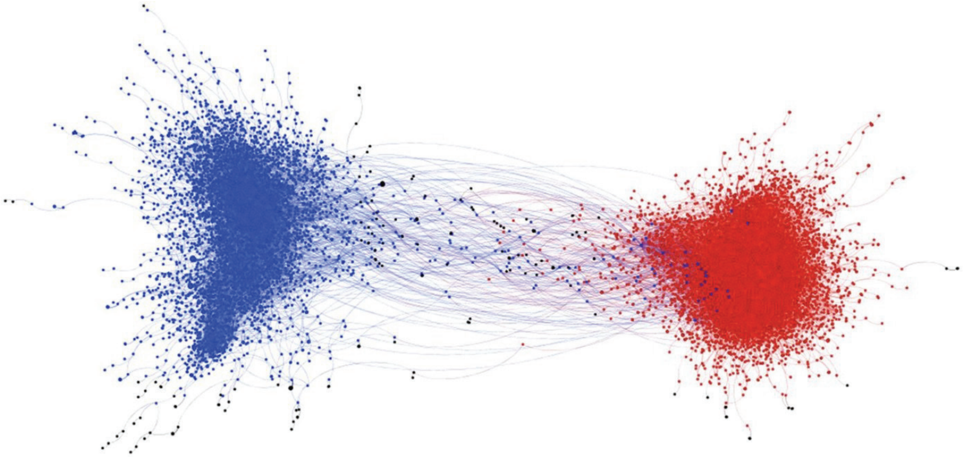

Figure 1

A network graph of moral contagion on Twitter shaded by political ideology (*blue* represents a liberal mean, *red* represents a conservative mean). The graph depicts messages containing moral-emotional language and their retweet activity across three political topics (gun control, same-sex marriage, and climate change). Nodes represent a user who sent a message, and edges represent a user retweeting another user's message. The two large communities were shaded based on the mean ideology of each respective community. Notably, tweets containing moral-emotional language were more likely to be shared within (rather than across) political groups. Figure adapted from Brady et al. (2017).

2005). Furthermore, morality has a strong emotional component and is tied to emotions like anger, disgust, and outrage (Rozin et al. 1999, Salerno & Peter-Hagene 2013). Indeed, people report greater support for political issues that elicit higher levels of moral emotions, like disgust or anger (Clifford 2019). In contrast to nonmoral emotions, moral emotions are highly interpersonal and elicited by social injustice and are thus considered prosocial emotions—that is, they are primarily motivated by a desire to defend the interests of others rather than one's own (Haidt 2003). For instance, moral outrage is characterized by a strong desire to shame or punish individuals who have behaved unjustly toward others, with the ultimate goal of promoting social fairness and equity.

Morality evolved as cooperative intuitions that facilitated group living among our ancestors (Axelrod & Hamilton 1981, Krebs 2008, Leimar & Hammerstein 2001, Petersen et al. 2012). As such, moral emotions can be conducive to social change and collective action: They help encourage group cooperation and discourage selfishness. For example, when people feel guilty about their own moral shortcomings, they are more likely to act prosocially to alleviate their negative feelings (Rothschild & Keefer 2017). Moral outrage can also drive people to punish those who commit moral transgressions like harming others or cheating (Balliet et al. 2011, Boyd et al. 2003, Xiao & Houser 2011). As such, moral outrage can produce destructive outcomes and even lead to violence (Fiske & Rai 2014). Thus, morality can be a force for positive social change as well as dangerous behavior.

SOCIAL CONFLICT AND POLARIZATION MANIFEST ONLINE

Throughout human history, conflicts over moral issues were resolved in face-to-face interactions. However, social interactions are increasingly moving into an online realm. With the advent of social media, the interactions among social groups increasingly take place in a fundamentally different environment than that of our ancestors (Li et al. 2018). Nonetheless, social media reproduces—and often exaggerates—the social behavior and intergroup dynamics that emerge from our group identities, moral sensibilities, and political polarization. To some extent, online spaces reflect offline priorities. Indeed, it is possible to infer people's personality traits and political affiliations based on their online social networks (Barberá 2015, Bond & Messing 2015). Moreover, people appear to display similar levels of political engagement (Quintelier & Theocharis 2013) and hostility (Bor & Petersen 2022) in online and offline contexts. However, people nevertheless perceive online discussions to be far more hostile (Bor & Petersen 2022). These perceptions of hostility may have significant implications for users, leading people to withdraw from social and political debates (Duggan 2017, Andresen et al. 2022) and possibly exacerbating real-world conflict (Mooijman et al. 2018).

The unique structure of the online environment can exacerbate certain adverse consequences of morality and politics (Brady et al. 2020a). As noted above, people encounter more moral transgressions on social media than in real life or in other forms of media such as television and print media (Crockett 2017). This may be due, in part, to the size of our online social networks. Offline, our social groups are limited in size, with an average of around 150 people (Hill & Dunbar 2003). With online platforms like Twitter, Facebook, TikTok, and Reddit, messages are instantly disseminated across networks, providing opportunities to connect with thousands of other people. Because people are increasingly using these platforms to disclose their beliefs and opinions, social media is loaded with moral content (Crockett 2017, McLoughlin et al. 2021). As such, people are regularly exposed to others' moral values and transgressions, including those of people with very different belief systems, resulting in an environment that is often morally charged or even hostile. Additionally, as social networks become larger, people's sense of morality becomes more generalized (e.g., varying on a single dimension of moral to immoral), as opposed to localized (e.g., varying across many dimensions and context dependent) (Jackson et al. 2023). Thus, people may adapt to the complexity of large online social networks by treating morality in a simpler way, identifying people or groups as good versus evil rather than as complex figures with a blend of positive and negative characteristics.

MORAL EMOTIONS CAPTURE ATTENTION ONLINE

Part of the reason that social media appears to inflame our moral beliefs and behavior is that we have a hard time ignoring moral content. Several large studies have found that social media posts that contain moral emotional language are more likely to be shared on social media than those that do not contain moral emotional language (Brady et al. 2017). According to a recent meta-analysis of 4,821,006 messages, each moral-emotional word added to a social media post is associated with 12% more retweets (Brady & Van Bavel 2021a). This pattern has been observed now across 27 different studies and is consistent across lay people and political elites (Brady et al. 2018). Similarly, news that is framed through a moral lens also receives more shares than nonmorally framed news (Valenzuela et al. 2017). This provides an incentive for users to employ moral emotional language to help spread their content and build a following (informally known as clout) on certain social media platforms.

Morality might be psychologically potent in the attention economy of social media because we are attuned to moral stimuli (Gantman & Van Bavel 2015). For instance, when people were asked to identify whether a string of letters presented rapidly (40 milliseconds) was a word, they were more likely to correctly recognize moral words as words compared to nonmoral words (Gantman & Van Bavel 2014). Moreover, people were especially likely to detect moral content when they were recently exposed to an injustice (Gantman & Van Bavel 2016). In turn, the moral words that capture attention are more likely to be shared on social media (Brady et al. 2020b). This suggests that the massive exposure to immoral events on social media (Crockett 2017) might make people especially attuned to other ambiguous moral content as they scroll through their newsfeeds.

REINFORCEMENT AND SOCIAL NORMS INCREASE MORAL OUTRAGE

Social norms are powerful drivers of behavior (Cialdini & Goldstein 2004). Social status hinges on adherence to societal norms, and people readily modify their behaviors to gain the approval of others (Mendes & Koslov 2013, Simpson et al. 2017). People will knowingly answer a question incorrectly to fit in (Asch et al. 1938), homeowners will decrease their power usage to match their neighbors (Schultz et al. 2007), and people will donate to charity if they believe it is normative (Alpizar et al. 2008, Smith et al. 2015). The speed and scale of social media messages can accelerate communication, conformity, and the enforcement of norms—including moral norms. For instance, social movements have used social media hashtags (e.g., #metoo, #blacklivesmatter) that could spread rapidly to huge audiences. While this can mobilize social action, it can also facilitate mass harassment campaigns, mob-like behavior, and conspiracy theories—known as morally motivated networked harassment (Marwick 2021). This networked harassment is a mechanism to enforce norm violations. In morally motivated networked harassment, a member of a social network or online community accuses a target of violating their network's norms, triggering moral outrage. Network members send harassing messages to the target, reinforcing their adherence to the norm and signaling group membership. Thus, social media can amplify norm following and enforcement at a speed and scale that would be difficult to implement in the real world.

People often perceive social media conversations to be more hostile than in-person interactions. However, social media may not cause us to act more aggressively, but rather exposes us to more aggressive individuals. Indeed, aggressive individuals appear to act equally aggressive in both offline and online contexts (Bor & Petersen 2022). Some researchers theorize that the absence of nonverbal cues that are essential in face-to-face communication may exaggerate the perceived intensity of users' outrage (Lieberman & Schroeder 2020). People often rely on nonverbal and auditory cues when trying to infer emotions (Hall & Schmid 2007, Kraus 2017, Zaki et al. 2009). However, many social media platforms primarily use text communication, which may lead users to overstate their outrage to compensate for the lack of nonverbal cues and clearly signal their membership in a particular group. Similarly, people tend to misinterpret social media posts as expressing greater outrage than the author actually feels (Brady et al. 2023). This can further inflate perceptions of online outrage around moral issues.

The anonymity of social media may also embolden some people to express views and emotions they may not disclose in face-to-face interactions (Nitschinsk et al. 2022, Zimmerman & Ybarra 2016). The lack of accountability of online environments can promote disinhibited and aggressive behavior, enabling people to use more divisive language online (Lieberman & Schroeder 2020, Suler 2004). For instance, people are less likely to use inflammatory language in online comments when identifying information is tagged to a comment (Cho & Kwon 2015). The discrepancy between anonymous and identifiable behavior demonstrates how social media serves as an accelerant for controversial content.

The relational mobility afforded by social media ensures that even those who burn bridges with their original social groups can find new groups that agree with their extreme opinions. People are easily able to move out of one group and into a new group where their reputations do not follow them due to the sheer number of online communities and increased anonymity. Moreover, motivated individuals can scale their harassment to numerous targets, and influencers can mobilize large groups in targeted harassment campaigns (Marwick 2021). These features can increase exposure to hostility even if the number of hostile individuals does not increase.

In the cacophony of online voices, moral outrage is attention grabbing and particularly effective at signaling group affiliation. For instance, outrage on a moral issue can signal an individual's social identity to like-minded others and boost their reputation as a trustworthy group member (Brady & Van Bavel 2021b, Marwick 2021). From an evolutionary standpoint, moral behaviors such as third-party punishment signal trustworthiness to the group and can therefore elevate status and achieve social recognition (Boyd & Richerson 1992, Johnen et al. 2018). Although this dynamic worked well for upholding moral norms within small communities (Boyd et al. 2003, Henrich et al. 2005, Rockenbach & Milinski 2006, Xiao & Houser 2011), social media has transformed moral discourse. By design, social media offers high-visibility and low-cost forms of engagement, such as comments and likes, which make it easy for users to punish, shame, and pile on transgressors. Social media may inflame these innate tendencies to punish wrongdoers, and compared to in-person punishment, online punishment comes with few drawbacks—it is more accessible and offers a large audience.

Although social media may facilitate outrage, norm enforcement, and public shaming, this can have unintended consequences for moral judgment. The larger the number of people who publicly criticize a moral transgression, the more likely third parties are to empathize with the original perpetrator of the transgression; this is known as "the paradox of viral outrage" (Sawaoka & Monin 2018). Instead of gaining social recognition for being morally good, large groups who sanction transgressors are then seen as bullies, and the person who committed the original transgression is given more sympathy than they would have been otherwise (Sawaoka & Monin 2018). In other words, social media may change the reputational dynamics in moral conflicts and lead to some unintended social consequences.

DESIGN FEATURES AND ALGORITHMS MAY AMPLIFY MORAL CONTENT AND CONFLICT

The design and incentives of social media accelerate moral content. For example, social media increases the visibility of moral conflicts: Hostile conversations appear to be more visible online (Bor & Petersen 2022). Moreover, they may be amplified by algorithms if they generate engagement. Thus, the incentive structure of online platforms is conducive to promoting intergroup conflict. Research suggests that dunking on out-group members is a strong predictor of engagement on social media, as divisive content is more likely to go viral (Rathje et al. 2021). Indeed, mentioning an out-group member is the strongest predictor of social media engagement out of a variety of predictors measured in an analysis of Facebook and Twitter posts ($N = 2{,}730{,}215$) by politicians and partisan news media sources (Rathje et al. 2021). Out-group language in particular predicts "angry" and "haha" reactions, indicating that mentions of the out-group may be closely linked to emotions such as anger, contempt, or mockery (see **Figure 2**). In other words, shaming or ridiculing political opponents appears to be a key driver of engagement on social media.

There is serious debate about whether social media platforms might foster echo chambers online, where people mostly hear perspectives from other in-group members (J. Haidt & C. Bail, unpublished manuscript). Some scholars argue that social media and the Internet expose people to more cross-cutting information and perspectives (Eady et al. 2019, Fletcher et al. 2021, Guess et al. 2021), while others argue that online, people are primarily exposed to politically congruent

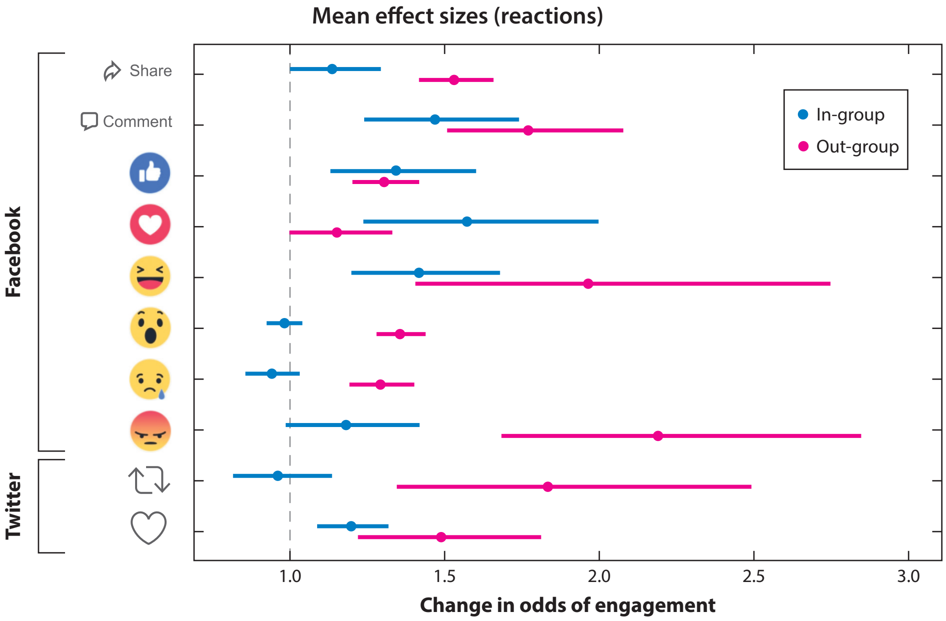

Figure 2

An analysis of Facebook and Twitter posts from politicians and news media sources found that mentions of the out-group strongly predicted social media engagement online. As shown above, out-group words (or words referring to the opposing political party) predicted Facebook shares, comments, "angry" reactions, and "haha" reactions as well as retweets. In-group words (or words referring to one's own political party) predicted "love" reactions and "like" reactions, but overall, in-group language received much lower engagement than out-group language online. These findings illustrate how social media might reward and amplify polarizing content. Figure adapted from Rathje et al. (2021).

information or people (Bakshy et al. 2015). This may reflect a tendency to prefer to make social ties with those who are close but also slightly more extreme in their opinions than ourselves—known as acrophily (Goldenberg et al. 2023). However, not all platforms are created equal, and some are more likely to foster echo chambers than others (Cinelli et al. 2021, Yarchi et al. 2021). One cross-platform analysis of homophily (the tendency for people to affiliate with those who are similar to them) and affective polarization (the tendency to view opposing partisans negatively and co-partisans positively) found that Twitter had the most homophily and Facebook had the lowest (Yarchi et al. 2021). This suggests that echo chambers may be due to specific design features, social norms, and algorithms on different platforms. This also underscores the need for more research on cross-platform differences in social media.

It is unclear how breaking down echo chambers might impact moral psychology. A long tradition of research on contact theory suggests that exposure to ethnically, morally, and politically diverse people and content might be associated with reduced polarization (Allport 1954, Paluck et al. 2019). However, simply exposing people to out-groups or politically incongruent information does not always have positive effects. For instance, one online experiment found that some people who are exposed to opposing partisan accounts on social media actually experience a backfire effect, such that they become more entrenched in their own beliefs (Bail et al. 2018). That said, this type of backfire effect might be the exception rather than the rule (Guess & Coppock 2020, Reinero et al. 2023, Wood & Porter 2019). As such, exposing people to counter-attitudinal information may not always be beneficial in every context, and this is especially true of social media contexts, where people are likely to be exposed to divisive viewpoints from their out-group.

As we noted above, one design feature of social media is that it enables people to express outrage online with little cost (Brady et al. 2020a). Users can easily show their support for (or disapproval of) issues through comments, shares, and likes—all of which are low-cost forms of engagement that signal affiliation. This feedback can then serve as a type of reinforcement learning that further fuels outrage (Brady et al. 2021). Because expressions of outrage earn a higher proportion of positive social feedback, this behavior is more likely to be rewarded. Users are encouraged to express even more outrage on social media, creating a self-perpetuating outrage cycle (Berger & Milkman 2012, Brady et al. 2017).

Social media algorithms further accelerate this cycle of outrage by promoting related ideological content. Repeated exposure to political content can amplify affective polarization and reinforce extremist views (Cho et al. 2020, Yarchi et al. 2021). As outrage on social media becomes more commonplace, people may express more outrage than they actually feel, as users internalize outrage expressions as the social norms of the community. Ideological extremists are less sensitive to reinforcement learning than typical users and participate in online outrage if it is considered a norm in their community (Brady et al. 2021). This may contribute to ideological extremists dominating online discussions.

MOTIVATIONS FOR EXPRESSING OUTRAGE

People express moral outrage online for a variety of reasons, such as to demonstrate their commitment to certain moral principles, signal their membership in their group, or enhance their reputation as moral individuals (Brady et al. 2020a). Although outrage could be characterized as a prosocial emotion (Haidt 2003), the expression of outrage can be motivated by both selfless and selfish goals. Outrage can indeed be expressed in the interest of others, for example, to bring attention to an injustice or mobilize collective action (Spring et al. 2018); but it can also be self-serving, allowing people to demonstrate their moral superiority, garner social approval, and acquire status (Tosi & Warmke 2020). In both cases, individuals must weigh the benefits of voicing outrage against the potential consequences, such as interpersonal or public backlash.

Outrage on social media can be propelled by intrinsic ideological goals, whereby individuals prioritize raising awareness about an issue over potential social or professional repercussions. In the lab, people behave morally due to an inherent desire to pursue justice, even if it is personally costly. For example, people are often willing to sacrifice monetary gains to punish a wrongdoer rather than earn money without administering punishment (Bernhard et al. 2006, Henrich et al. 2005). Further, people donate to those in need even if there is no public recognition of their contribution, demonstrating an intrinsic motivation to help others (Kraus & Callaghan 2016, Sisco & Weber 2019). On social media, however, it can be difficult to disentangle whether users' outrage reflects genuine altruism or is a disingenuous attempt to boost their social image.

Regardless of intent, moral outrage can encourage an active response, such as protests for social justice (Hutcherson & Gross 2011, Jost et al. 2018, Spring et al. 2018, van Zomeren et al. 2004). People may therefore want to capitalize on the public nature of social media, which enables their messages to reach a broad audience (Carr 2017). Online outrage can raise awareness about an issue, motivate people to fight against perceived injustice, and advocate for social change (Van de Vyver & Abrams 2015). Collective expressions of outrage create a social understanding that certain behaviors are wrong and can bind people together through social movements (Shteynberg et al. 2017, Spring et al. 2018). This is why a wide variety of social activism takes place on social media.

However, public expressions of outrage can lead to benefits for individuals, and this behavior online may therefore be motivated by extrinsic goals. Humans are drawn to seek social status and will often behave differently in social settings to boost their status and demonstrate their trustworthiness (Cheng et al. 2013, Nesi & Prinstein 2015). For instance, when given the opportunity to punish wrongdoers, people administer harsher punishments in the presence of others than in private settings (Kurzban et al. 2007). Given that moral behavior drives social evaluation, people may voice moral judgments on social media to demonstrate their own moral commitment or undermine a rival's moral beliefs, garnering a higher online status through a behavior known as moral grandstanding (Grubbs et al. 2019). Moral grandstanding refers to the pursuit of social status through moral reputation.

Status seeking via outrage expression is effective because people often equate moral outrage expressions with the speaker's own morality (Skitka 2010, Tosi & Warmke 2016). When people express outrage, they are perceived as having greater conviction in their moral values, thus earning them a higher moral status in the eyes of others. Moral grandstanding is considered a form of free riding in which individuals join in on others' moral discourse, benefiting from the issue without making any meaningful or costly contributions (Tosi & Warmke 2016). People engage in moral grandstanding by inflating and moralizing minor or symbolic issues, which provides an unambiguous public signal about their commitment to a moral code and then incites others to join in. This can generate a flurry of online outrage that vastly exceeds the original transgression.

This appeared to be at play in January 2021, when a father took to Twitter in an attempt to humorously document his daughter's struggle to use a can opener (Di Placido 2021). He jokingly recounted her frustration as he encouraged her to figure out how to use it on her own. Commenters, however, were livid and accused the author—who became widely known as Bean Dad—of child abuse. Although the post was intended to portray a funny story, people were quick to extract a moral issue and condemn the author until he deleted his account and publicly apologized. The reward structure of Twitter amplified this incident to a prominent trending news story that eventually overshadowed breaking news about Donald Trump attempting to overturn the results of the 2020 US presidential election in Georgia. This type of disproportional moralized collective response is commonplace on social media platforms, as users are primed to interpret content through a moral lens and are rewarded for participating in moral discourse. For instance, users who comment on trending topics, hashtags, or threads are promoted by the platforms and receive additional attention, engagement, and followers.

INDIVIDUAL DIFFERENCES IN ONLINE MORALITY

People who engage in aggressive and hostile behaviors in the offline world are also likely to do so in the online world, where their actions are accelerated by the structure of the Internet (Bor & Petersen 2022; Kowalski et al. 2014, 2019). For instance, people with disagreeable traits are more likely to be online cyberbullies (Van Geel et al. 2017). A meta-analysis of antisocial online behaviors found that psychopathy was the strongest and most consistent predictor of antisocial online behaviors, followed by Machiavellianism and everyday sadism (Moor & Anderson 2019). These findings clarify the psychology of trolls who derail social media for their own amusement (Buckels et al. 2014). Furthermore, Machiavellianism, narcissism, and psychopathy (known collectively as the Dark Triad of personality) are associated with virtuous victim signaling—a form of status seeking in which people portray themselves as a trustworthy, moral character, which garners them sympathy and elevates their moral status (Ok et al. 2021). Another strategy to acquire status is through dominance—the use of force and intimidation to induce fear—which is related to a range of aggressive behaviors (Cheng et al. 2013). Studies suggest that feelings of marginalization coupled with dominance orientation make people more likely to engage in hostility on social media (Bor & Petersen 2022, Petersen et al. 2023), and thus social media might provide a powerful tool for achieving these motivations. In this way, individual differences shape the interplay between social media and morality (Crosier et al. 2012).

While personality traits are important predictors of aggressive behaviors in moralized discussions on social media, there is good reason to believe that people engage in this manner to further their political causes. Online extremists participate in a significant amount of online discourse, and this suggests that a large proportion of posts may be from individuals with antisocial tendencies. Moreover, antidemocratic politicians, Republicans, and foreign operatives were the most likely to employ polarizing rhetoric online in recent years (Simchon et al. 2022, Valentino-DeVries & Eder 2022). Political events in the real world often trigger reactions in social media, as hate speech and misinformation often increase during elections and political turmoil (Kim 2022, Rasmussen & Petersen 2022, Siegel et al. 2021).

At the individual level, politics also motivates online behavior (e.g., Erjavec & Kovačič 2012, Rasmussen 2023). One study of fake news sharing on Twitter found that partisan polarization is the primary psychological motivation behind fake news sharing, and thus it "is fueled by the same psychological motivations that drive other forms of partisan behavior, including sharing partisan news from traditional and credible news sources" (Osmundsen et al. 2021, p. 999). This suggests that political motivations are key drivers of content consumption and sharing on social media, regardless of the content's veracity (Robertson et al. 2023b). Similarly, frequent commenters on Facebook are more politically interested, have more polarized opinions, and, in turn, use more toxic language (Kim et al. 2021). Thus, while people who hold certain personality traits are more likely to engage in aggressive activities in general, moralized political discussions on social media might amplify, reward, and select for people with certain traits.

EXTREMISTS DOMINATE ONLINE CONVERSATIONS

Moral discourse online is dominated by ideological extremists. Political extremists, both liberal and conservative, show higher levels of outrage on social media compared to politically moderate users (Brady et al. 2021). Given that outrage is attention grabbing and social media algorithms prioritize popular content, our online feeds are saturated with expressions of outrage distorting our perceptions of polarization. In fact, heavy social media users are prone to false consensus effects, meaning they overestimate how widely held their views are (Bunker & Varnum 2021). People also tend to overestimate the degree of polarization (Levendusky & Malhotra 2016, Westfall et al. 2015) and animosity (Mernyk et al. 2022, Ruggeri et al. 2021) between political groups, which may be exacerbated by the abundance of divisive content they see online. This also feeds into political segregation on social media. Because there is such a large volume of people to connect with in the online context, there are significantly more extreme voices to affiliate with than there were in prior social groups that were limited by geography (Petersen et al. 2020, Törnberg 2022).

Extremists' participation online may have a polarizing effect on peoples' more moderate ideological views. Extremists tend to be dogmatic and to present their beliefs as facts (Harris & Van Bavel 2020, Toner et al. 2013), and they may incite others to become more ideologically polarized. In other words, extremists' contribution on social media may inaccurately depict political groups as more polarized than they truly are (known as false polarization), yet their participation may lead to real ramifications in which they convince others of their extreme views. In discussions with like-minded individuals, people can be persuaded to adopt more extreme ideological positions, leaving an interaction with more polarized views than they initially had (Mackie & Cooper 1984, Moscovici & Zavalloni 1969). This push toward extremist views may also result in social media acting as a catalyst for conspiratorial thinking. Social media use is associated with greater conspiratorial beliefs among people who are predisposed to conspiratorial thinking (Enders et al. 2023). This may lead to conspiracy believers to self-select into online communities that in turn share and reinforce their beliefs (Robertson et al. 2022).

Social media might also cause people to encounter more divisive and polarizing content than they actually want to see. For example, a recent study explored differences in the type of content people think social media amplifies versus the type of content people would want social media platforms to amplify in an ideal world (Rathje et al. 2023). In line with some prior research, a representative sample of Americans reported thinking that moral outrage, polarizing content, misinformation, and outgroup animosity all tend to go viral on social media platforms. However, when asked what type of content they would want to go viral on social media, people reported that they do not want to see divisive content and instead would like more accurate, thoughtful, or nuanced content to go viral. In other words, even though people engage with divisive content on social media—perhaps because it captures attention and drives engagement—this behavior does not align with people's stated preferences (see **Figure 3**). Interestingly, both Democrats and Republicans reported that divisive content should be much less viral than it currently is, indicating that there is broad consensus across the political spectrum that people do not want divisive content to go viral.

CONSEQUENCES

What are the consequences of the rapid spread of moralized content online? Whistleblower Frances Haugen, a former Facebook project manager, testified to the US Congress that the social media site's products "harm children, stoke division and weaken our democracy" and that congressional intervention is needed (McCluskey 2021). While Haugen and many scholars have argued that various societal ills are caused by social media (Fisher 2022b, Haidt 2022), social media can lead to many positive outcomes as well. Social media, like many other major innovations, such as the printing press, the telephone, and the television, has rapidly changed how people consume and share information. As with these earlier technological innovations, some argue there is a moral panic surrounding the consequences of the Internet and social media that strongly resembles prior moral panics (Orben 2020). Therefore, it is important to balance

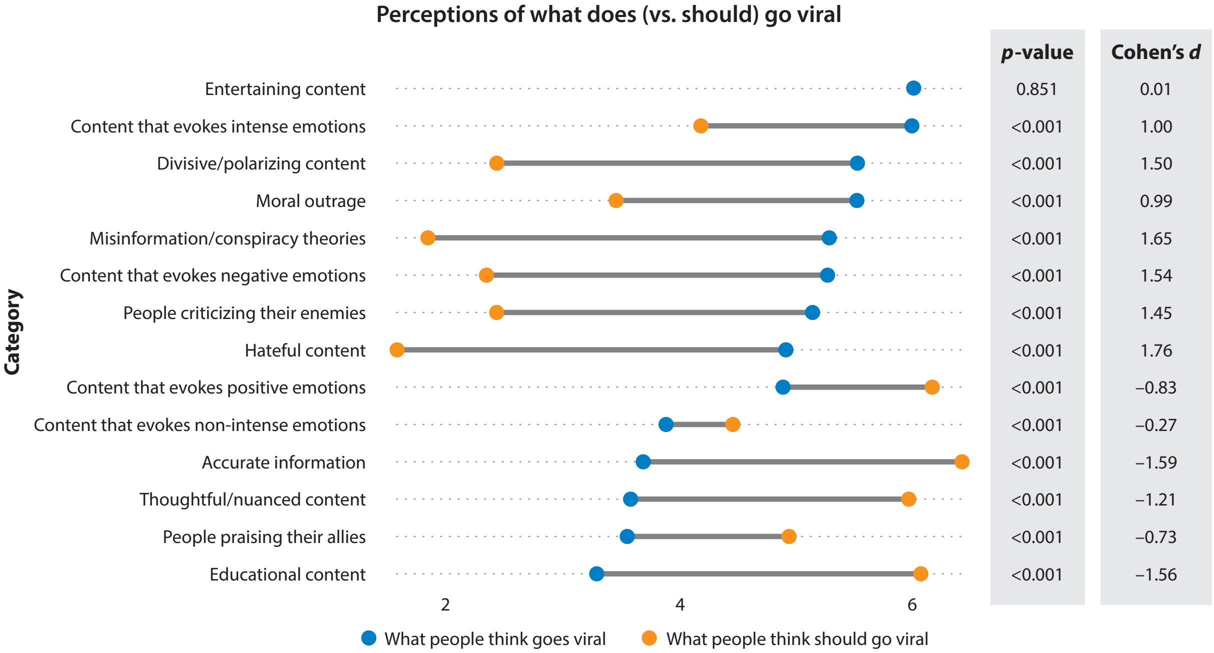

Figure 3

A representative sample of 511 US participants revealed that while people think that divisive content, moral outrage, misinformation, and people criticizing their enemies all go viral online, they report that they do not want these types of content to go viral. Instead, they would prefer that positive, thoughtful, or nuanced content went viral. Figure adapted from Rathje et al. (2023).

the moral trade-offs of any technology and develop practices and regulations that maximize the benefits of new technology and mitigate harm. Here, we review some of the consequences of social media, with a particular focus on the consequences of the spread of moral outrage on social media. These include both positive consequences, such as increased information spread, awareness, and political action, and negative consequences, such as polarization, misinformation spread, public health risks, and political violence.

INFORMATION SHARING AND AWARENESS

A major reason social media is different from prior forms of media, such as books, newspapers, and television, is that information can be shared with much lower effort, with a greater number of people, and much faster than ever before—as there are fewer gatekeepers for the sharing of online information (and misinformation). Open access to online platforms can be beneficial, enabling marginalized voices to face lower barriers to disseminate moral messages and find social support (Schmitz et al. 2022). This lack of gatekeepers can facilitate organizing and collective action (Spring et al. 2018). For instance, social media helped mobilize massive global protests movements around sexual harassment (e.g., #metoo) and racial discrimination (e.g., #blacklivesmatter). However, this can also have unintended implications. For example, social media might be making nonviolent protests less effective because they are easier to organize without leadership, which makes them less enduring. The result is massive rallies that emerge on the streets overnight but often fizzle just as quickly (Fisher 2022a). Thus, the long-term implications of organizing on social media remain unclear.

The lack of barriers and editorial oversight can also be harmful. Anyone has the opportunity to create a social media account and post content that rapidly goes viral even if that content consists of baseless conspiracy theories or misinformation (Vosoughi et al. 2018). This creates a paradox: Social media can be a liberating technology, giving marginalized voices who are normally excluded the chance to speak, but it can also be a repressive technology, providing bad actors, conspiracy theorists, or propagandists with the opportunity to spread harmful messages (Tucker et al. 2017). This makes it extremely difficult to evaluate the net moral impact of social media on society (see Brady & Crockett 2019).

Social media may also accelerate certain political processes, both good and bad. Social media access has been implicated in increasing access to vital information for individuals in authoritarian regimes, but also in increasing support for populist parties and spurring social conflict. For instance, one study found that the rollout of 3G Internet access around the world was associated with decreased support for political institutions and increased support for populist political parties (Guriev et al. 2021). Part of this trend might be explained by the Internet's informing users about government corruption (Zhuravskaya et al. 2020).

Some have also argued that social media may be particularly beneficial to those in developing democracies and authoritarian regimes (as opposed to established democracies) because of increased information access that would otherwise be suppressed by the state (Lorenz-Spreen et al. 2023). While social media can provide useful information, it can also lead to the spread of harmful misinformation. For instance, social media may have played a role in increasing violent sectarian conflicts in Myanmar and Sri Lanka through the spreading of harmful misinformation, or in the United States through the spreading of antidemocratic #stopthesteal conspiracy theories about the 2020 US presidential election (Fisher 2022b).

Information spread online is governed, in part, by social media algorithms, which might dramatically change the kind of information that people consume (Rathje et al. 2022b). Algorithms seem to prioritize the spread of information that is emotional, is polarizing, or captures our attention (Brady et al. 2020a,b). Thus, while more voices can be heard through social media, more extreme voices may be prioritized (Bail 2022) and propaganda can foster divisive content (Simchon et al. 2022), which might distort the information people consume. The lack of oversight can lead to the spread of misinformation that can undercut democratic institutions and might even lead to ethnic or political violence.

POLARIZATION

Another potential consequence of the spread of moral and emotional content online is increased polarization (Iyengar et al. 2019). A number of scholars have expressed concern that social media might be fueling political polarization (Haidt 2022, Harris et al. 2023, Kubin & von Sikorski 2021, Van Bavel et al. 2021b), and this notion seems to be supported by some empirical evidence. For instance, one randomized controlled trial found that deactivating Facebook for one month reduced ideological polarization (or polarization around political issues) and marginally reduced affective polarization in the United States (Allcott et al. 2020). However, social media may have different effects on polarization in non-US contexts. For example, deleting Facebook during genocide remembrance week in Bosnia actually increased ethnic polarization, especially if people had particularly homogeneous offline social networks (Asimovic et al. 2021). Presumably, these individuals were in an echo chamber in their real life, and social media gave them exposure to alternative perspectives about historical atrocities. Thus, social media might have very different impacts on polarization in different places, demonstrating a need for more global research on the effects of social media.

The increased spread of moralized content online may further fuel polarization and intergroup conflict. For instance, one study found that moral-emotional words (like "hate" and "blame") were more likely to spread within groups (i.e., in echo chambers) than across party lines (Brady et al. 2017; see **Figure 1**). When people do not use moral-emotional language to discuss the same political issues, there is little evidence of polarization along partisan or ideological lines. Another study found that, while someone's expression of moral-emotional language made in-group members view that person as a good group member, it made out-group members view that person as less open-minded and more partisan (Brady & Van Bavel 2021b). Other work finds that moralization may amplify a preference to share both true and false politically congruent partisan news, which can amplify polarization (Marie et al. 2023). Moreover, using moral-emotional rhetoric can trigger hate speech among recipients: Each additional moral word is associated with a 11–16% increase in eliciting hate speech from readers (Solovev & Pröllochs 2023). Thus, moralized rhetoric may increase polarization, harassment, misinformation, and false norms about the amount of outrage and hostility present in people's online social networks.

MISINFORMATION AND CONSPIRACY THEORIES

Social media allows for the rapid spread of conspiracy theories, which are often motivated by moral concerns. Because social media platforms allow information to spread with less friction than in traditional news outlets, such as newspapers or television news, misinformation and conspiracy theories are able to go rapidly viral (Robertson et al. 2022, Van Bavel et al. 2021a, van der Linden et al. 2021). One study found that fact-checked false stories spread faster and farther than fact-checked true stories on Twitter (Vosoughi et al. 2018). This was especially true of emotional stories on political topics, which were presumably loaded with moral content that drove people to spread them.

Many studies have found that misinformation and moral outrage are often closely intertwined (Marie et al. 2023, Vosoughi et al. 2018). One analysis found that COVID-19 rumors that contained moral-emotional language were more likely to spread widely on Twitter (Solovev & Pröllochs 2022). Another analysis of Facebook and Twitter posts found that the misinformation containing moral outrage (as detected by a moral outrage machine-learning classifier) was more likely to spread than misinformation that did not contain moral outrage (McLoughlin et al. 2021). As we noted above, posts with misinformation tend to contain high-arousal emotions such as surprise or disgust (Vosoughi et al. 2018) and often derogate out-group members (Osmundsen et al. 2021). These features may help explain why misinformation spreads so quickly within social networks. Thus, algorithms that promote outrage may prioritize promoting false information and conspiracy theories.

There is debate about the size of the online misinformation problem, with some scholars arguing that the misinformation discourse is plagued by alarmist narratives (Altay et al. 2023). While many might think that exposure to misinformation is very common, exposure to untrustworthy news sites is actually rare, and people tend to see far more legitimate news online (Altay et al. 2022). Additionally, true news is more likely to generate engagement than false news on Reddit (Bond & Garrett 2023). However, other research suggests that these studies massively underestimate the scope and spread of misinformation because they are largely limited to text. For instance, a full 23% of political images on Facebook contain misinformation (Yang et al. 2023). These contradicting findings may reflect difficulties in defining misinformation. If misinformation is defined as news websites that frequently publish fabricated news stories, then misinformation exposure is rare. However, if misinformation is defined as misleading claims made by politicians and public figures (Mosleh & Rand 2022) or political images (Yang et al. 2023), then misinformation exposure is likely much more common and potentially impactful. For this reason, it might be more useful to focus on the impact of misinformation on important domains (e.g., vaccine hesitancy, false beliefs about election outcomes) rather than focusing solely on the volume of misinformation.

THE CONSEQUENCES OF ONLINE MORALITY AND MISINFORMATION

Social media and morality may have a number of substantial consequences for public health and democratic functioning. Early in the COVID-19 pandemic, attitudes and behavior related to the pandemic were especially polarized, with Republicans and Democrats in the United States having strikingly different attitudes about public health behaviors and the COVID-19 vaccine (Gollwitzer et al. 2020, Van Bavel et al. 2023). Social media may be responsible in part for this polarization around public health beliefs. For instance, polarized rhetoric about the pandemic on Twitter was highly correlated with polarized beliefs about the risks of the pandemic several days later (Simchon et al. 2022). Americans with anti-vaccine and pro-vaccine attitudes clustered into distinct echo chambers on Twitter (Rathje et al. 2022). Pro-vaccine individuals in the United States tended to follow liberal politicians on Twitter (such as Kamala Harris and Hillary Clinton), whereas antivaccine individuals were more likely to follow conservative politicians and influencers (such as Candace Owens or Tucker Carlson). However, these echo chambers about vaccine attitudes were not observed in the United Kingdom, suggesting that political leaders in the United States may have polarized the pandemic (Rathje et al. 2022a; see also Altay et al. 2022). Other research suggests that misinformation spread on Twitter is linked to lower county-level vaccine uptake (Pierri et al. 2022), and people who got their news primarily from Facebook were even more vaccine hesitant than those who got their news primarily from Fox News (Lazer et al. 2021). Other research has found a causal link between exposure to vaccine misinformation and reduced intentions to get vaccinated (Loomba et al. 2021).

The spread of moral messages on social media can lead to important collective action (such as protests, political participation, and mobilization), as well as harmful behaviors (such as political violence). One perspective is that that outrage online is a critical force for collective action (Spring et al. 2018). Some groups have harnessed shared outrage to effect change by spreading awareness through hashtags or organizing protests. One analysis during the 2013–2014 Maidan protests in Ukraine found that protestors who took to Facebook spread awareness, garnering millions of likes (Jost et al. 2018; see **Figure 4**). Similarly, the #MeToo hashtag was included in over 13 million tweets and was highly correlated with real-life social and political events (Williams et al. 2019). As such, there is clear evidence that social media is used to facilitate collective action to benefit marginalized groups; however, the impact of these efforts is often unclear.

Others argue that online outrage is unlikely to lead to effective collective action (Brady & Crockett 2019). The flood of outrage on these sites is often directed at different issues and may ultimately dilute collective efforts. The convenience of social media may therefore backfire and draw attention away from social issues. Moreover, moralized rhetoric online may spillover into real-world violence (Fisher 2022b, Mooijman et al. 2018). For instance, Facebook was implicated in an outbreak of ethnic cleansing in Myanmar against the Rohingya (Fisher 2022b). Moreover, the amount of moral convergence in online groups (e.g., people in a group sharing the same moral views) predicts radicalization (Atari et al. 2022), and, as we mentioned above, tweets containing moral-emotional language are more likely to receive hate speech in the replies (Solovev & Pröllochs 2023). In sum, while moral outrage may lead to collective action, the research literature so far has found stronger association with more extreme or harmful action such as violence or radicalization.

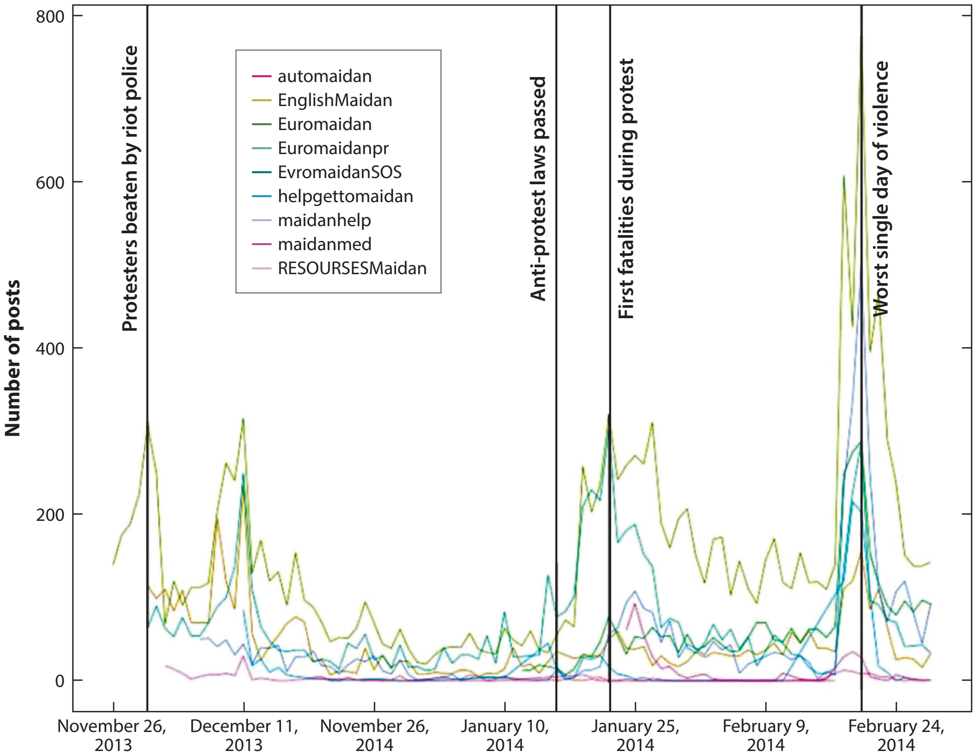

Figure 4

Number of Facebook posts related to the Maidan protests in Ukraine by Facebook group (automaidan, EnglishMaidan, EuroMaidan, euromaidanpr, EvromaidanSOS, helpgettomaidan, maidanhelp, maidanmed, RESOURSESMaidan), date, and protest-related events (2013–2014). This figure shows the number of Facebook posts on each of several key pages related to the protests over time. Vertical lines mark key protest-related events in order to highlight the ways in which activity spiked on certain pages in response to offline political developments. Figure adapted from Jost et al. (2018).

DISCUSSION AND FUTURE DIRECTIONS

While people's morality is shaped by stable factors, the interplay between social media and morality sometimes accelerates the content people consume and how they express themselves on social media. This can include factual as well as fictitious or conspiratorial content. Given that people are unlikely to change their behaviors and social media is designed to reinforce online outrage, we outline a range of measures that can ameliorate some of the adverse consequences of the interplay between social media and morality.

SOCIAL MEDIA ACCELERATES MORAL DYNAMICS

Social media seems to amplify exposure to moral content and action. In the political domain, there is ample evidence that affective polarization is on the rise (Finkel et al. 2020, Iyengar & Westwood 2015), leading to concerns about reduced support for democratic norms (Broockman et al. 2022, Kingzette et al. 2021). This tendency may be playing out online as well, as recent studies suggest that in political discussions online, divisive or moralized content is more likely to go viral (Brady et al. 2017, Rathje et al. 2021), which may fuel intergroup conflict both offline and online (Van

Bavel et al. 2022). We do not believe that social media is primarily responsible for polarization and threats to democracy, but simply that it might serve as an accelerant of existing cultural divisions.

As a result of exposure to divisive content, people may have distorted perceptions of those who are different from themselves. People hold exaggerated meta-perceptions of their political outgroups and overstate the magnitude of partisan animosity the political out-group feel about one's own political in-group (Ruggeri et al. 2021), and they also overestimate the degree of polarization between groups. There is reason to believe that social media exacerbates these issues by disproportionately exposing people to extremists. These false beliefs help explain why social media can fuel (moral) conflicts, as most people are more moderate than people think (Levendusky & Malhotra 2016) and consume news from the same sources on social media (Guess 2021), while those who engage on social media are the most engaged partisans (Settle 2018). How social media algorithms should be modified is not exclusively an empirical question. Yet examining how social media and morality affect societies—for better or worse—requires access to researchers and transparency from the platforms.

ADDRESSING THE INTERPLAY BETWEEN SOCIAL MEDIA AND MORALITY

There is a growing recognition among scholars and the public that social media has deleterious consequences for society, and there is a growing appetite for greater transparency and some form of regulation of social media platforms (Rathje et al. 2023). To address the adverse consequences of social media, solutions at the system level are necessary (e.g., Chater & Loewenstein 2022), but individual- or group-level solutions may be useful for creating behavioral change before system-level change is in place and for increasing public support for system-level solutions (Koppel et al. 2023). We discuss a range of solutions that address the adverse consequences of the interplay between social media and morality.

Regulation is one of the most heavily debated ways of mitigating the adverse features of social media. Regulating social media can be done both on platforms and at the national or cross-national level, but it always involves discussions about who should decide what should be allowed on which platforms (Kaye 2019). Currently, there is relatively little editorial oversight on the content even on mainstream platforms, yet the association of regulation with censorship makes regulation inherently controversial. For instance, Americans believe that social media companies censor political viewpoints (Vogels et al. 2020) and believe it is hard to regulate social media because people cannot agree upon what should and should not be removed (Pew Res. Cent. 2019). Moreover, authoritarian states can suppress dissent through the regulation of speech on social media.

In general, people on the political left are more supportive of regulating social media platforms (Kozyreva et al. 2023, Rasmussen 2022), reflecting liberals' general tendency to be more supportive, and conservatives' tendency to more opposing, of regulatory policies (e.g., Grossman & Hopkins 2016). In the context of social media content, one explanation is that left-leaning people infer more harm from aggressive behaviors. In other words, they may perceive immoral behaviors on social media as more harmful for the victim, which in turn justifies regulation (Crawford 2017, Graham et al. 2009, Walter & Redlawsk 2019).

Political allegiances also shape perceptions of hostility (Muddiman 2017, Mutz 2015) and censorship (Amira et al. 2021, Ashokkumar et al. 2020, Lelkes & Westwood 2017), and they may make people willing to compromise civil liberties for advancing political goals (Frederiksen 2022, Graham & Svolik 2020, Simonovits et al. 2022, Svolik 2020). However, new research suggests there is more consensus in public opinion on regulating social media than previously thought. One study found that Americans overestimate how much members of the opposing party "cancel," or publicly shame, others (Dias et al. 2022). People across the political spectrum believe that misinformation and hate speech should be restricted based on severity, regardless of the target or topic (Kozyreva et al. 2023, Rasmussen 2022, Rathje et al. 2023). These studies suggest that regulating threatening content that causes severe harm is in line with public opinion, although there may be disagreement on what constitutes severe harm.

A range of interventions have been developed to mitigate the adverse consequences of the interplay between social media and morality such as misinformation (Guess et al. 2020a, Pennycook et al. 2021, Rasmussen et al. 2022, van der Linden et al. 2020), polarization (Hartman et al. 2022), and hate speech (Yildirim et al. 2021), and they are already being deployed and tested online (see Van Bavel et al. 2021a). One promising strategy is motivating bystanders to denounce hostility or harassment on social media (Munger 2017). While most people are passive on social media and only consume content (Settle 2018), encouraging users to participate in online discourse and to react negatively to hostility may improve the social media environment. Decades of research in social psychology highlight how people are attentive to social norms put forth by one's in-group. Thus, harnessing people's social identities or social norms can motivate prosocial behavior (Van Bavel & Packer 2021). For instance, identity- and empathy-based bystander interventions on social media can effectively reduce incivility (Munger 2020), hate speech (Hangartner et al. 2021, Munger 2017, Siegel & Badaan 2020), and misinformation (Pretus et al. 2022). This suggests that peer influence may be effective in shaping norms in social media environments when the peers are credible members of one's in-group (Paluck & Green 2009, Paluck et al. 2021).

Other solutions focus on shifting incentive structures to encourage people to share more constructive or accurate content. Certain features of the social media algorithms can be adjusted to encourage the spread of more productive content. For example, social media platforms put increased weight on algorithmic inputs that are associated with more productive content such as "love" or "like" reactions, which were more strongly associated with in-group positivity than out-group negativity (Rathje et al. 2021). Other experimental work has found that increasing people's motivation to be accurate, and reducing people's motivation to view content through a partisan lens, improves the accuracy of people's beliefs and sharing decisions (Rathje et al. 2022a). As such, aspects of social media design, incentive structures, and algorithms can all be shifted to encourage people to share more positive content at scale.

Addressing the adverse consequences of the interplay between social media and morality through regulation, interventions, or incentive structures on social media might serve as an effective amelioration strategy. However, these strategies often mitigate, rather than address, the root of the problem. In other words, morality and social identities shape behavior on social media and people share misinformation because it supports their social, political, and moral goals. Thus, regulation, incentive structures, and interventions may help mitigate the negative consequences of social media in the short-term, but they would likely be more effective if they addressed the broader social and cultural problems that underlie problematic behaviors. As such, it is critical to foster the necessary public support for addressing the roots of conflicts in the offline world so that they do not spillover into social media (Chater & Loewenstein 2022, Koppel et al. 2023). Moreover, any substantial interventions on platforms should engage a broad body of stakeholders and maximize transparency and accountability.

FUTURE RESEARCH

With established social media platforms regularly changing their designs and new sites capturing our attention, research on social media is constantly evolving. As these platforms have achieved broad global penetration, there is an urgent need for more cross-cultural research on social media.

The large majority of published research has focused on North American and European issues and samples. For instance, 86% of the research on social media and polarization is from the United States (Kubin & von Sikorski 2021). This is despite the fact that 9 out of 10 countries with the highest numbers of Facebook users are located in the Global South (Ghai et al. 2022). Moreover, recent research suggests that the psychological impact of social media appears to be very different in the Global North versus the Global South (Ghai et al. 2023), or in established democracies versus less-established democracies (Lorenz-Spreen et al. 2023). As Tucker et al. (2017, p. 48) observe, "social media can be at once a technology of liberation, a technology useful to authoritarian governments bent on stifling dissent, and a technology for empowering those seeking to challenge the status quo in democratic societies." Therefore, our understanding of the consequences of moral discourse on social media may not accurately reflect the experience of users from other parts of the world or understudied communities. Testing theories in distinct contexts is crucial to advance our understanding of social media and morality, as some processes might lead to distinct consequences under different circumstances (Asimovic et al. 2021).

Another challenge is that researchers must continually update their conceptions of social media, including how distinct platforms may serve different purposes. While studies on Twitter are relatively abundant, other popular platforms such as Reddit, LinkedIn, Instagram, or YouTube have not received the same level of attention (Iandoli et al. 2021). As such, our ability to generalize across platforms is severely limited. This is in part due to the fact that Twitter data have been historically much easier and less expensive to access, whereas platforms such as Facebook, Instagram, or TikTok share very little data with researchers. Providing data access to researchers is crucial to enhance our understanding of how different social media platforms, design choices, and algorithms influence human behavior (Tucker et al. 2017). Overall, social media platforms are very hesitant to share data and are not transparent about how their underlying algorithms work, even though 92% of people support social media companies being more transparent about how their algorithms work (Rathje et al. 2022b; see **Figure 3**). Hopefully, research access and tools will improve to allow more rigorous comparisons of the role of morality across platforms.

Second, experimental field research is key to understanding the consequences of social media at the individual and societal level (Mosleh et al. 2022; cf. Munger 2019 for a critique). Previous field experiments have examined whether online news sources and social media increase polarization (Bail et al. 2018) and how negative emotionality (Robertson et al. 2023a) and toxicity affect engagement (Beknazar-Yuzbashev et al. 2022). Far more work is needed to connect these online patterns of behavior to real world outcomes. Field experiments can be embedded in panel surveys with nationally representative samples to assess the causal effects of interventions over time (Guess et al. 2021), and measures of behavior and beliefs can be connected to online data to identify important relationships (Rathje et al. 2022). Thus, field research and experiments on social media provide an opportunity to assess causal hypotheses in ecologically valid settings.

Interdisciplinary work is also necessary to fully understand how social media impacts individuals, groups, and societies. In recent years, the social sciences provided early guidance on responses to the COVID-19 pandemic to inform policy decisions (Ruggeri et al. 2022, Van Bavel et al. 2020). In the same vein, psychologists should continue to inform policy decisions related to social media by integrating knowledge from within the social sciences (e.g., political science, psychology, communication, sociology, media studies) and beyond (e.g., computer science, data science, engineering, neuroscience). Bringing together scholars from these disciplines will also allow for better methods and theoretical frameworks. Of course, this type of research requires the necessary funding and institutional support to bring together experts from different backgrounds. We hope to see more interdisciplinary grant opportunities, centers, and conferences to help spawn the next generation of research on social media and morality.

CONCLUSION

Human morality plays a central role in the social and attentional dynamics of social media. Social media is projected to grow to nearly 6 billion users in a few years, and millions of young people are now on these platforms much of the time. This means that the relationship between our paleolithic moral emotions and this godlike technology is only bound to grow in importance. Understanding how our morality shapes our behavior on social media and how social media, in turn, shapes our moral psychology and behavior is likely to be a critical topic for scholars and policy makers for the foreseeable future. Given the immense scale of the topic, understanding these bidirectional influences will likely prove essential to building more humane technology and fostering healthier societies.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

Allcott H, Braghieri L, Eichmeyer S, Gentzkow M. 2020. The welfare effects of social media. *Am. Econ. Rev.* 110(3):629–76

Allport GW. 1954. *The Nature of Prejudice*. Boston: Addison-Wesley

Alpizar F, Carlsson F, Johansson-Stenman O. 2008. Anonymity, reciprocity, and conformity: evidence from voluntary contributions to a national park in Costa Rica. *J. Public Econ.* 92(5):1047–60

Altay S, Berriche M, Acerbi A. 2023. Misinformation on misinformation: conceptual and methodological challenges. *Soc. Media Soc.* 9(1). **https://doi.org/10.1177/20563051221150412**

Altay S, Kleis Nielsen R, Fletcher R. 2022. Quantifying the "infodemic": People turned to trustworthy news outlets during the 2020 coronavirus pandemic. *J. Quant. Descr. Digit. Media* 2. **https://doi.org/10.51685/jqd.2022.020**

Amira K, Wright JC, Goya-Tocchetto D. 2021. In-group love versus out-group hate: Which is more important to partisans and when? *Political Behav.* 43:473–94

Andresen MJ, Karg STS, Rasmussen SHR, Pradella L, Rasmussen J, et al. 2022. *Danskernes oplevelse af had på sociale medier* [The Danes' experience of hate on social media]. Rep., Aarhus Univ., Aarhus, Den. **https://pure.au.dk/portal/files/271291115/Danskernes_oplevelse_af_had_pa_de_sociale_medier_Rapport_Aarhus_Universitet_.pdf**

Aramovich NP, Lytle BL, Skitka LJ. 2012. Opposing torture: moral conviction and resistance to majority influence. *Soc. Influence* 7(1):21–34

Asch SE, Block H, Hertzman M. 1938. Studies in the principles of judgments and attitudes: I. Two basic principles of judgment. *J. Psychol.* 5(2):219–51

Ashokkumar A, Talaifar S, Fraser WT, Landabur R, Buhrmester M, et al. 2020. Censoring political opposition online: Who does it and why. *J. Exp. Soc. Psychol.* 91:104031

Asimovic N, Nagler J, Bonneau R, Tucker JA. 2021. Testing the effects of Facebook usage in an ethnically polarized setting. *PNAS* 118(25):e2022819118

Atari M, Davani AM, Kogon D, Kennedy B, Ani Saxena N, et al. 2022. Morally homogeneous networks and radicalism. *Soc. Psychol. Pers. Sci.* 13(6):999–1009

Axelrod R, Hamilton WD. 1981. The evolution of cooperation. *Science* 211(4489):1390–96

Bail C. 2022. *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing*. Princeton, NJ: Princeton Univ. Press

Bail CA, Argyle LP, Brown TW, Bumpus JP, Chen H, et al. 2018. Exposure to opposing views on social media can increase political polarization. *PNAS* 115(37):9216–21

Bak-Coleman JB, Alfano M, Barfuss W, Bergstrom CT, Centeno MA, et al. 2021. Stewardship of global collective behavior. *PNAS* 118(27):e2025764118

Bakshy E, Messing S, Adamic LA. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348(6239):1130–32

Balliet D, Mulder LB, Van Lange PAM. 2011. Reward, punishment, and cooperation: a meta-analysis. *Psychol. Bull.* 137:594–615

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

Beknazar-Yuzbashev G, Jiménez Durán R, McCrosky J, Stalinski M. 2022. *Toxic content and user engagement on social media: evidence from a field experiment.* Work. Pap., Univ. Chicago, Chicago

Berger J, Milkman KL. 2012. What makes online content viral? *J. Mark. Res.* 49(2):192–205

Bernhard H, Fischbacher U, Fehr E. 2006. Parochial altruism in humans. *Nature* 442(7105):912–15

Boehm C. 2012. *Moral Origins: The Evolution of Virtue, Altruism, and Shame*. New York: Soft Skull Press

Bond RM, Garrett RK. 2023. Engagement with fact-checked posts on Reddit. *PNAS Nexus* 2(3):pgad018

Bond RM, Messing S. 2015. Quantifying social media's political space: estimating ideology from publicly revealed preferences on Facebook. *Am. Political Sci. Rev.* 109(1):62–78

Bor A, Petersen MB. 2022. The psychology of online political hostility: a comprehensive, cross-national test of the mismatch hypothesis. *Am. Political Sci. Rev.* 116(1):1–18

Boyd R, Richerson PJ. 1992. Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethol. Sociobiol.* 13(3):171–95

Boyd R, Gintis H, Bowles S, Richerson PJ. 2003. The evolution of altruistic punishment. *PNAS* 100(6):3531–35

Brady WJ, Crockett MJ. 2019. How effective is online outrage? *Trends Cogn. Sci.* 23:79-80

Brady WJ, Crockett MJ, Van Bavel JJ. 2020a. The MAD model of moral contagion: the role of motivation, attention, and design in the spread of moralized content online. *Perspect. Psychol. Sci.* 15(4):978–1010

Brady WJ, Gantman AP, Van Bavel JJ. 2020b. Attentional capture helps explain why moral and emotional content go viral. *J. Exp. Psychol. Gen.* 149(4):746–56

Brady WJ, McLoughlin K, Doan TN, Crockett MJ. 2021. How social learning amplifies moral outrage expression in online social networks. *Sci. Adv.* 7(33):eabe5641

Brady WJ, McLoughlin KL, Torres MP, Luo KF, Gendron M, Crockett MJ. 2023. Overperception of moral outrage in online social networks inflates beliefs about intergroup hostility. *Nat. Hum. Behav.* 7(6):917–27

Brady WJ, Van Bavel JJ. 2021a. *Estimating the effect size of moral contagion in online networks: a pre-registered replication and meta-analysis*. Work. Pap., Northwestern Univ., Evanston, IL

Brady WJ, Van Bavel JJ. 2021b. *Social identity shapes antecedents and functional outcomes of moral emotion expression in online networks*. Work. Pap., Northwestern Univ., Evanston, IL

Brady WJ, Wills JA, Burkart D, Jost JT, Van Bavel JJ. 2018. An ideological asymmetry in the diffusion of moralized content on social media among political leaders. *J. Exp. Psychol. Gen.* 148(10):1802–13

Brady WJ, Wills JA, Jost JT, Tucker JA, Bavel JJV. 2017. Emotion shapes the diffusion of moralized content in social networks. *PNAS* 114(28):7313–18

Broockman DE, Kalla JL, Westwood SJ. 2022. Does affective polarization undermine democratic norms or accountability? Maybe not. *Am. J. Political Sci.* **https://doi.org/10.1111/ajps.12719**

Buckels EE, Trapnell PD, Paulhus DL. 2014. Trolls just want to have fun. *Pers. Individ. Differ.* 67:97–102

Bunker CJ, Varnum MEW. 2021. How strong is the association between social media use and false consensus? *Comput. Hum. Behav.* 125:106947

Carr CT. 2017. Social media and intergroup communication. In *The Oxford Encyclopedia of Intergroup Communication*, Vol. 2, ed. H Giles, J Harwood, pp. 349–67. Oxford, UK: Oxford Univ. Press

Carr CT, Hayes RA. 2015. Social media: defining, developing, and divining. *Atl. J. Commun.* 23(1):46–65

Chater N, Loewenstein G. 2022. The i-frame and the s-frame: how focusing on individual-level solutions has led behavioral public policy astray. *Behav. Brain Sci.* In press. **https://doi.org/10.1017/S0140525X22002023**

Cheng JT, Tracy JL, Foulsham T, Kingstone A, Henrich J. 2013. Two ways to the top: evidence that dominance and prestige are distinct yet viable avenues to social rank and influence. *J. Pers. Soc. Psychol.* 104(1):103–25

Cho D, Kwon KH. 2015. The impacts of identity verification and disclosure of social cues on flaming in online user comments. *Comput. Hum. Behav.* 51:363–72

Cho J, Ahmed S, Hilbert M, Liu B, Luu J. 2020. Do search algorithms endanger democracy? An experimental investigation of algorithm effects on political polarization. *J. Broadcast. Electron.* 64(2):150–72

Cialdini RB, Goldstein NJ. 2004. Social influence: compliance and conformity. *Annu. Rev. Psychol.* 55:591–621

Cinelli M, De Francisci Morales G, Galeazzi A, Quattrociocchi W, Starnini M. 2021. The echo chamber effect on social media. *PNAS* 118(9):e2023301118

Clifford S. 2019. How emotional frames moralize and polarize political attitudes. *Political Psychol.* 40(1):75–91

Cohen AB. 2015. Religion's profound influences on psychology: morality, intergroup relations, self-construal, and enculturation. *Curr. Dir. Psychol. Sci.* 24(1):77–82

Crawford JT. 2017. Are conservatives more sensitive to threat than liberals? It depends on how we define threat and conservatism. *Soc. Cogn.* 35(4):354–73

Crockett MJ. 2017. Moral outrage in the digital age. *Nat. Hum. Behav.* 1(11):769–71

Crosier BS, Webster GD, Dillon HM. 2012. Wired to connect: evolutionary psychology and social networks. *Rev. Gen. Psychol.* 16(2):230–39

Di Placido D. 2021. The ballad of "Bean Dad" shows the cruel, petty side of Twitter. *Forbes*, Jan. 5. **https://www.forbes.com/sites/danidiplacido/2021/01/05/the-ballad-of-bean-dad-shows-the-cruel-petty-side-of-twitter/?sh=72128800648e**

Dias NC, Druckman JN, Levendusky M. 2022. *How and why Americans misperceive the prevalence of, and motives behind, "cancel culture."* Work. Pap., Univ. Pa., Philadelphia

Duggan M. 2017. Online harassment 2017. *Pew Research Center*, July 11. **https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/**

Eady G, Nagler J, Guess A, Zilinsky J, Tucker JA. 2019. How many people live in political bubbles on social media? Evidence from linked survey and Twitter data. *Sage Open* 9(1):2158244019832705

Ellemers N, van den Bos K. 2012. Morality in groups: on the social-regulatory functions of right and wrong. *Soc Pers. Psychol. Compass* 6(12):878–89

Enders AM, Uscinski JE, Seelig MI, Klofstad CA, Wuchty S, et al. 2023. The relationship between social media use and beliefs in conspiracy theories and misinformation. *Political Behav*. 45:781–804

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

Fincher KM, Tetlock PE. 2016. Perceptual dehumanization of faces is activated by norm violations and facilitates norm enforcement. *J. Exp. Psychol. Gen.* 145(2):131–46

Finkel EJ, Bail CA, Cikara M, Ditto PH, Iyengar S, et al. 2020. Political sectarianism in America. *Science* 370(6516):533–36

Fisher M. 2022a. Even as Iranians rise up, protests worldwide are failing at record rates. *New York Times*, Sept. 30. **http://www.nytimes.com/2022/09/30/world/middleeast/iran-protests-haiti-russia-china.html**

Fisher M. 2022b. *The Chaos Machine: The Inside Story of How Social Media Rewired Our Minds and Our World*. New York: Little, Brown & Co.

Fiske AP, Rai TS. 2014. *Virtuous violence: Hurting and Killing to Create, Sustain, End, and Honor Social Relationships*. Cambridge, UK: Cambridge Univ. Press

Fletcher R, Kalogeropoulos A, Nielsen RK. 2021. More diverse, more politically varied: how social media, search engines and aggregators shape news repertoires in the United Kingdom. *New Media Soc.* In press. **https://doi.org/10.1177/14614448211027393**

Frederiksen KVS. 2022. Does competence make citizens tolerate undemocratic behavior? *Am. Political Sci. Rev.* 116(3):1147–53

Gantman AP, Van Bavel JJ. 2014. The moral pop-out effect: enhanced perceptual awareness of morally relevant stimuli. *Cognition* 132(1):22–29

Gantman AP, Van Bavel JJ. 2015. Moral perception. *Trends Cogn. Sci.* 19(11):631–33

Gantman AP, Van Bavel JJ. 2016. Exposure to justice diminishes moral perception. *J. Exp. Psychol. Gen.* 145(12):1728–39

Ghai S, Fassi L, Awadh F, Orben A. 2023. Lack of sample diversity in research on adolescent depression and social media use: a scoping review and meta-analysis. *Clin. Psychol. Sci.* in press. **https://doi.org/10.1177/21677026221114859**

Ghai S, Magis-Weinberg L, Stoilova M, Livingstone S, Orben A. 2022. Social media and adolescent well-being in the Global South. *Curr. Opin. Psychol.* 46:101318

Goldenberg A, Abruzzo JM, Huang Z, Schöne J, Bailey D, et al. 2023. Homophily and acrophily as drivers of political segregation. *Nat. Hum. Behav.* 7:219–30

Gollwitzer A, Martel C, Brady WJ, Pärnamets P, Freedman IG, Knowles ED, Van Bavel JJ. 2020. Partisan differences in physical distancing are linked to health outcomes during the COVID-19 pandemic. *Nat. Hum. Behav.* 4:1186–97

Graham J, Haidt J, Nosek BA. 2009. Liberals and conservatives rely on different sets of moral foundations. *J. Pers. Soc. Psychol.* 96(5):1029–46

Graham MH, Svolik MW. 2020. Democracy in America? Partisanship, polarization, and the robustness of support for democracy in the United States. *Am. Political Sci. Rev.* 114(2):392–409

Grossman M, Hopkins DA. 2016. *Asymmetric Politics: Ideological Republicans and Group Interest Democrats*. New York: Oxford Univ. Press

Grubbs JB, Warmke B, Tosi J, James AS, Campbell WK. 2019. Moral grandstanding in public discourse: status-seeking motives as a potential explanatory mechanism in predicting conflict. *PLOS ONE* 14(10):e0223749

Guess AM. 2021. (Almost) everything in moderation: new evidence on Americans' online media diets. *Am. J. Political Sci.* 65(4):1007–22

Guess AM, Barberá P, Munzert S, Yang J. 2021. The consequences of online partisan media. *PNAS* 118(14):e2013464118

Guess AM, Coppock A. 2020. Does counter-attitudinal information cause backlash? Results from three large survey experiments. *Br. J. Political Sci.* 50(4):1497–515

Guess AM, Lerner M, Lyons B, Montgomery JM, Nyhan B, Reifler J, Sircar N. 2020a. A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *PNAS* 117(27):15536–45

Guess AM, Nyhan B, Reifler J. 2020b. Exposure to untrustworthy websites in the 2016 US election. *Nat. Hum. Behav.* 4(5):472–80

Guriev S, Melnikov N, Zhuravskaya E. 2021. 3G internet and confidence in government. *Q. J. Econ.* 136(4):2533–613

Haidt J. 2003. The moral emotions. In *Handbook of Affective Sciences*, ed. RJ Davidson, KR Scherer, HH Goldsmith, pp. 852–70. New York: Oxford Univ. Press

Haidt J. 2007. The new synthesis in moral psychology. *Science* 316(5827):998–1002

Haidt J. 2022. Why the past 10 years of American life have been uniquely stupid. *Atlantic*, April 11. **https://www.theatlantic.com/magazine/archive/2022/05/social-media-democracy-trust-babel/629369/**

Haidt J, Graham J. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Soc. Justice Res.* 20(1):98–116

Hall JA, Schmid Mast M. 2007. Sources of accuracy in the empathic accuracy paradigm. *Emotion* 7(2):438–46

Hangartner D, Gennaro G, Alasiri S, Bahrich N, Bornhoft A, et al. 2021. Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *PNAS* 118(50):e2116310118

Harris EA, Rathje S, Robertson C, Van Bavel JJ. 2023. The SPIR Model of Social Media and Polarization: Exploring the Role of Selection, Platform Design, Incentives, and Real-World Context. *Int. J. Commun.* In press

Harris EA, Van Bavel JJ. 2020. Pre-registered replication of "Feeling superior is a bipartisan issue: Extremity (not direction) of political views predicts belief superiority." *Psychol. Sci.* 32(3):451–58

Hartman R, Blakey W, Womick J, Bail C, Finkel EJ, et al. 2022. Interventions to reduce partisan animosity. *Nat. Hum. Behav.* 6(9):1194–205

Harv. Mag. 2009. An intellectual entente. *Harvard Magazine*, Sept. 10. **https://www.harvardmagazine.com/breaking-news/james-watson-edward-o-wilson-intellectual-entente**

Henrich J, Boyd R, Bowles S, Camerer C, Fehr E, et al. 2005. "Economic man" in cross-cultural perspective: behavioral experiments in 15 small-scale societies. *Behav. Brain. Sci.* 28(6):795–815

Hill RA, Dunbar RIM. 2003. Social network size in humans. *Hum. Nat.* 14(1):53–72

Hornsey MJ, Majkut L, Terry DJ, McKimmie BM. 2003. On being loud and proud: non-conformity and counter-conformity to group norms. *Br. J. Soc. Psychol.* 42(3):319–35

Hutcherson CA, Gross JJ. 2011. The moral emotions: a social-functionalist account of anger, disgust, and contempt. *J. Pers. Soc. Psychol.* 100(4):719–37

Iandoli L, Primario S, Zollo G. 2021. The impact of group polarization on the quality of online debate in social media: a systematic literature review. *Technol. Forecast. Soc. Change* 170:120924

Iyengar S, Lelkes Y, Levendusky M, Malhotra N, Westwood SJ. 2019. The origins and consequences of affective polarization in the United States. *Annu. Rev. Political Sci.* 22:129–46

Iyengar S, Westwood SJ. 2015. Fear and loathing across party lines: new evidence on group polarization. *Am. J. Political Sci.* 59(3):690–707

Jackson JC, Halberstadt J, Takezawa M, Kongmeng L, Smith KM, et al. 2023. Generalized morality culturally evolves as an adaptive heuristic in large social networks. PsyArXiv, March 22. **https://doi.org/10.31234/osf.io/sx4rt**

Jarudi I, Kreps T, Bloom P. 2008. Is a refrigerator good or evil? The moral evaluation of everyday objects. *Soc. Justice Res.* 21:457–69

Johnen M, Jungblut M, Ziegele M. 2018. The digital outcry: What incites participation behavior in an online firestorm? *New Media Soc.* 20(9):3140–60

Jost JT, Barberá P, Bonneau R, Langer M, Metzger M, et al. 2018. How social media facilitates political protest: information, motivation, and social networks. *Adv. Political Psychol.* 39:85–118

Kaye D. 2019 August 12. Four questions about regulating online hate speech. *OneZero*, Aug. 12. **https://onezero.medium.com/four-questions-about-online-hate-speech-ae3e0a134472**

Kim JW, Guess A, Nyhan B, Reifler J. 2021. The distorting prism of social media: how self-selection and exposure to incivility fuel online comment toxicity. *J. Commun.* 71(6):922–46

Kim T. 2022. Violent political rhetoric on Twitter. *Political Sci. Res. Methods*. In press. **https://doi.org/10.1017/psrm.2022.12**

Kingzette J, Druckman JN, Klar S, Krupnikov Y, Levendusky M, Ryan JB. 2021. How affective polarization undermines support for democratic norms. *Public Opin. Q.* 85(2):663–77

Koppel L, Robertson CE, Doell KC, Javeed AM, Rasmussen J, et al. 2023. Individual-level solutions may support system-level change—if they are internalized as part of one's social identity. *Brain Behav. Sci.* In press

Kowalski RM, Giumetti GW, Schroeder AN, Lattanner MR. 2014. Bullying in the digital age: a critical review and meta-analysis of cyberbullying research among youth. *Psychol. Bull.* 140(4):1073–137

Kowalski RM, Limber SP, McCord A. 2019. A developmental approach to cyberbullying: prevalence and protective factors. *Aggress. Violent Behav.* 45:20–32

Kozyreva A, Herzog SM, Lewandowsky S, Hertwig R, Lorenz-Spreen P, et al. 2023. Resolving content moderation dilemmas between free speech and harmful misinformation. *PNAS* 120(7):e2210666120

Kraus MW. 2017. Voice-only communication enhances empathic accuracy. *Am. Psychol.* 72(7):644–54

Kraus MW, Callaghan B. 2016. Social class and prosocial behavior: the moderating role of public versus private contexts. *Soc. Psychol. Pers. Sci.* 7(8):769–77

Krebs DL. 2008. Morality: an evolutionary account. *Perspect. Psychol. Sci.* 3(3):149–72

Kross E, Verduyn P, Sheppes G, Costello CK, Jonides J, Ybarra O. 2021. Social media and well-being: pitfalls, progress, and next steps. *Trends Cogn. Sci.* 25(1):55–66

Kubin E, von Sikorski C. 2021. The role of (social) media in political polarization: a systematic review. *Ann. Int. Commun. Assoc.* 45(3):188–206

Kurzban R, DeScioli P, O'Brien E. 2007. Audience effects on moralistic punishment. *Evol. Hum. Behav.* 28(2):75–84

Lazer D, Green J, Ognyanova K, Baum M, Lin J, et al. 2021. *The COVID States Project #57: social media news consumption and COVID-19 vaccination rates*. Rep., COVID States Proj. **https://www.covidstates.org/reports/social-media-news-consumption-and-covid-19-vaccination-rates**

Leimar O, Hammerstein P. 2001. Evolution of cooperation through indirect reciprocity. *Proc. R. Soc. B* 268(1468):745–53

Lelkes Y, Westwood SJ. 2017. The limits of partisan prejudice. *J. Politics* 79(2):485–501

Levendusky MS, Malhotra N. 2016. (Mis) perceptions of partisan polarization in the American public. *Public Opin. Q.* 80(S1):378–91

Li NP, van Vugt M, Colarelli SM. 2018. The evolutionary mismatch hypothesis: implications for psychological science. *Curr. Dir. Psychol. Sci.* 27(1):38–44

Lieberman A, Schroeder J. 2020. Two social lives: how differences between online and offline interaction influence social outcomes. *Curr. Opin. Psychol.* 31:16–21

Loomba S, de Figueiredo A, Piatek SJ, de Graaf K, Larson HJ. 2021. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nat. Hum. Behav.* 5(3):337–48

Lorenz-Spreen P, Oswald L, Lewandowsky S, Hertwig R. 2023. A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nat. Hum. Behav.* 7:74–101

Mackie D, Cooper J. 1984. Attitude polarization: effects of group membership. *J. Pers. Soc. Psychol.* 46(3):575–85

Marie A, Altay S, Strickland. 2023. Moralization and extremism robustly amplify myside sharing. *PNAS Nexus* 2:pgad078

Marwick AE. 2021. Morally motivated networked harassment as normative reinforcement. *Soc. Media Soc.* 7(2). **https://doi.org/10.1177/20563051211021378**

McCluskey M. 2021. 4 big takeaways from the Facebook whistleblower congressional hearing. *TIME*, Oct. 5. **https://www.time.com/6104070/facebook-whistleblower-congressional-hearing-takeaways/**

McLoughlin KL, Brady WJ, Crockett MJ. 2021. The role of moral outrage in the spread of misinformation. In *Technology, Mind & Society 2021 Conference Proceedings*. Washington, DC: Am. Psychol. Assoc. **https://doi.org/10.1037/tms0000136**

Mendes WB, Koslov K. 2013. Brittle smiles: positive biases toward stigmatized and outgroup targets. *J. Exp. Psychol. Gen.* 142(3):923–33

Mernyk JS, Pink SL, Druckman JN, Willer R. 2022. Correcting inaccurate metaperceptions reduces Americans' support for partisan violence. *PNAS* 119(16):e2116851119

Mooijman M, Hoover J, Lin Y, Ji H, Dehghani M. 2018. Moralization in social networks and the emergence of violence during protests. *Nat. Hum. Behav.* 2(6):389–96

Moor L, Anderson JR. 2019. A systematic literature review of the relationship between dark personality traits and antisocial online behaviours. *Pers. Individ. Differ.* 144:40–55

Morant L. 2018. The truth behind 6 second ads. *Medium*, Febr. 8. **https://medium.com/@Lyndon/the-tyranny-of-six-seconds-592b94160877**

Moscovici S, Zavalloni M. 1969. The group as a polarizer of attitudes. *J. Pers. Soc. Psychol.* 12(2):125–35

Mosleh M, Pennycook G, Rand DG. 2022. Field experiments on social media. *Curr. Dir. Psychol. Sci.* 31(1):69–75

Mosleh M, Rand DG. 2022. Measuring exposure to misinformation from political elites on Twitter. *Nat. Commun.* 13(1):7144

Muddiman A. 2017. Personal and public levels of political incivility. *Int. J. Commun.* 11:21

Munger K. 2017. Tweetment effects on the tweeted: experimentally reducing racist harassment. *Political Behav.* 39:629–49

Munger K. 2019. The limited value of non-replicable field experiments in contexts with low temporal validity. *Soc. Media Soc.* 5(3). **https://doi.org/10.1177/2056305119859294**

Munger K. 2020. Don't @ me: experimentally reducing partisan incivility on Twitter. *J. Exp. Political Sci.* 8(2):102–16

Mutz DC. 2015. *In-Your-Face Politics: The Consequences of Uncivil Media*. Princeton, NJ: Princeton Univ. Press

Nesi J, Prinstein MJ. 2015. Using social media for social comparison and feedback-seeking: Gender and popularity moderate associations with depressive symptoms. *J. Abnorm. Child Psychol.* 43(8):1427–38

Nitschinsk L, Tobin SJ, Vanman EJ. 2022. The disinhibiting effects of anonymity increase online trolling. *Cyberpsychol. Behav. Soc. Netw.* 25(6):377–83

Ok E, Qian Y, Strejcek B, Aquino K. 2021. Signaling virtuous victimhood as indicators of Dark Triad personalities. *J. Pers. Soc. Psychol.* 120(6):1634–61

Orben A. 2020. The Sisyphean cycle of technology panics. *Perspect. Psychol. Sci.* 15(5):1143–57

Orben A, Przybylski AK. 2019. The association between adolescent well-being and digital technology use. *Nat. Hum. Behav.* 3(2):173–82

Orben A, Przybylski AK, Blakemore SJ, Kievit RA. 2022. Windows of developmental sensitivity to social media. *Nat. Commun.* 13(1):1649

Osmundsen M, Bor A, Vahlstrup PB, Bechmann A, Petersen MB. 2021. Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter. *Am. Political Sci. Rev.* 115(3):999–1015

Paluck EL, Green DP. 2009. Prejudice reduction: What works? A review and assessment of research and practice. *Annu. Rev. Psychol.* 60:339–67

Paluck EL, Green SA, Green DP. 2019. The contact hypothesis re-evaluated. *Behav. Public Policy* 3(2):129–58

Paluck EL, Porat R, Clark CS, Green DP. 2021. Prejudice reduction: progress and challenges. *Annu. Rev. Psychol.* 72:533–60

Pennycook G, Epstein Z, Mosleh M, Arechar AA, Eckles D, Rand DG. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature* 592(7855):590–95

Petersen M, Osmundsen M, Arceneaux K. 2023. The "need for chaos" and motivations to share hostile political rumors. *Am. Polit. Sci. Rev.* In press. **https://doi:10.1017/S0003055422001447**

Petersen M, Osmundsen M, Tooby J. 2020. The evolutionary psychology of conflict and the functions of falsehood. PsyArXiv, Aug. 29. **https://doi.org/10.31234/osf.io/kaby9**

Petersen MB, Sell A, Tooby J, Cosmides L. 2012. To punish or repair? Evolutionary psychology and lay intuitions about modern criminal justice. *Evol. Hum. Behav.* 33(6):682–95

Pew Res. Cent. 2019. The challenge of knowing what's offensive. *Pew Research Center*, June 19. **https://www.pewresearch.org/politics/2019/06/19/the-challenge-of-knowing-whats-offensive/**

Pierri F, Perry BL, DeVerna MR, Yang K-C, Flammini A, et al. 2022. Online misinformation is linked to early COVID-19 vaccination hesitancy and refusal. *Sci. Rep.* 12:5966

Pretus C, Javeed A, Hughes DR, Hackenburg K, Tsakiris M, et al. 2022. *The misleading count: an identity-based intervention to mitigate the spread of partisan misinformation*. Work. Pap., Univ. Auton. Barcelona, Barcelona, Spain

Quintelier E, Theocharis Y. 2013. Online political engagement, Facebook, and personality traits. *Soc. Sci. Comput. Rev.* 31(3):280–90

Rajkumar K, Saint-Jacques G, Bojinov I, Brynjolfsson E, Aral S. 2022. A causal test of the strength of weak ties. *Science* 377(6612):1304–10

Rasmussen J. 2022. When do the public support hate speech restrictions? Symmetries and asymmetries across partisans in Denmark and the United States. PsyArXiv, June 8. **https://doi.org/10.31234/osf.io/j4nuc**

Rasmussen J. 2023. Pathways to online political hostility on social media. PsyArXiv, March 27. **https://doi.org/10.31234/osf.io/r3y5u**

Rasmussen J, Lindekilde L, Petersen MB. 2022. Public health communication reduces COVID-19 misinformation sharing and boosts self-efficacy. PsyArXiv, July 7. **https://doi.org/10.31234/osf.io/8wdfp**

Rasmussen SHR, Petersen MB. 2022. From echo chambers to resonance chambers: how offline political events enter and are amplified in online networks. PsyArXiv, May 26. **https://doi.org/10.31234/osf.io/vzu4q**

Rathje S, He JK, Roozenbeek J, Van Bavel JJ, van der Linden S. 2022. Social media behavior is associated with vaccine hesitancy. *PNAS Nexus* 1(4):pgac207

Rathje S, Robertson C, Brady W, Van Bavel JJ. 2023. People think that social media platforms do (but should not) amplify divisive content. *Perspect. Psychol. Sci.* In press

Rathje S, Van Bavel JJ, van der Linden S. 2021. Out-group animosity drives engagement on social media. *PNAS* 118(26):e2024292118

Reinero DA, Harris EA, Rathje S, Duke A, Van Bavel JJ. 2023. Partisans are more likely to entrench their beliefs in misinformation when political outgroup members fact-check claims. PsyArXiv, May 11. **https://doi.org/10.31234/osf.io/z4df3**

Robertson CE, Pretus C, Rathje S, Harris EA, Van Bavel JJ. 2022. How social identity shapes conspiratorial belief. *Curr. Opin. Psychol.* 47:101423

Robertson CE, Pröllochs N, Schwarzenegger K, Pärnamets P, Van Bavel JJ, Feuerriegel S. 2023a. Negativity drives online news consumption. *Nat. Hum. Behav.* 7:812–22

Robertson RE, Green J, Ruck DJ, Ognyanova K, Wilson C, Lazer D. 2023b. Users choose to engage with more partisan news than they are exposed to on Google Search. *Nature* 618:342–48

Rockenbach B, Milinski M. 2006. The efficient interaction of indirect reciprocity and costly punishment. *Nature* 444:718–23

Rothschild ZK, Keefer LA. 2017. A cleansing fire: Moral outrage alleviates guilt and buffers threats to one's moral identity. *Motiv. Emot.* 41:209–29

Rozin P. 1999. The process of moralization. *Psychol. Sci.* 10(3):218–21

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). *J. Pers. Soc. Psychol.* 76(4):574–86

Ruggeri K, Stock F, Haslam SA, Capraro V, Boggio P, et al. 2022. Evaluating expectations from social and behavioral science about COVID-19 and lessons for the next pandemic. PsyArXiv, Oct. 10. **https://doi.org/10.31234/osf.io/58udn**

Ruggeri K, Većkalov B, Bojanić L, Andersen TL, Ashcroft-Jones S, et al. 2021. The general fault in our fault lines. *Nat. Hum. Behav.* 5(10):1369–80

Salerno JM, Peter-Hagene LC. 2013. The interactive effect of anger and disgust on moral outrage and judgments. *Psychol. Sci.* 24(10):2069–78

Sawaoka T, Monin B. 2018. The paradox of viral outrage. *Psychol. Sci.* 29(10):1665–78

Schmitz RM, Coley JS, Thomas C, Ramirez A. 2022. The cyber power of marginalized identities: intersectional strategies of online LGBTQ+ Latinx activism. *Fem. Media Stud.* 22(2):271–90

Schultz PW, Nolan JM, Cialdini RB, Goldstein NJ, Griskevicius V. 2007. The constructive, destructive, and reconstructive power of social norms. *Psychol. Sci.* 18(5):429–34

Settle JE. 2018. *Frenemies: How Social Media Polarizes America*. Cambridge, UK: Cambridge Univ. Press

Shteynberg G, Gelfand M, Imai L, Mayer DM, Bell C. 2017. Prosocial thinkers and the social transmission of justice. *Eur. J. Soc. Psychol.* 47(4):429–42

Siegel AA, Badaan V. 2020. # No2Sectarianism: experimental approaches to reducing sectarian hate speech online. *Am. Political Sci. Rev.* 114(3):837–55

Siegel AA, Nikitin E, Barberá P, Sterling J, Pullen B, et al. 2021. Trumping hate on Twitter? Online hate speech in the 2016 US election campaign and its aftermath. *Q. J. Political Sci.* 16(1):71–104

Simchon A, Brady WJ, Van Bavel JJ. 2022. Troll and divide: the language of online polarization. *PNAS Nexus* 1(1):pgac019

Simon B, Klandermans B. 2001. Politicized collective identity: a social psychological analysis. *Am. Psychol.* 56(4):319–31

Simonovits G, McCoy J, Littvay L. 2022. Democratic hypocrisy and out-group threat: explaining citizen support for democratic erosion. *J. Politics* 84(3):1806–11

Simpson B, Willer R, Harrell A. 2017. The enforcement of moral boundaries promotes cooperation and prosocial behavior in groups. *Sci. Rep.* 7:42844

Sisco MR, Weber EU. 2019. Examining charitable giving in real-world online donations. *Nat. Commun.* 10(1):3968

Skitka LJ. 2010. The psychology of moral conviction. *Soc. Pers. Psychol. Compass* 4(4):267–81

Skitka LJ, Bauman CW, Sargis EG. 2005. Moral conviction: another contributor to attitude strength or something more? *J. Pers. Soc. Psychol.* 88(6):895–917

Skitka LJ, Morgan GS. 2014. The social and political implications of moral conviction. *Adv. Political Psychol.* 35:95–110

Skitka LJ, Washburn AN. 2016. Are conservatives from Mars and liberals from Venus? Maybe not so much. In *Social Psychology of Political Polarization*, ed. P Valdesolo, J Graham, pp. 78–101. London: Routledge

Smith LGE, Thomas EF, McGarty C. 2015. "We must be the change we want to see in the world": integrating norms and identities through social interaction. *Political Psychol.* 36(5):543–57

Solovev K, Pröllochs N. 2022. Moral emotions shape the virality of COVID-19 misinformation on social media. ArXiv:2202.03590 [cs.SI]

Solovev K, Pröllochs N. 2023. Moralized language predicts hate speech on social media. *PNAS Nexus* 2:pgac281

Spring VL, Cameron CD, Cikara M. 2018. The upside of outrage. *Trends Cogn. Sci.* 22(12):1067–69

Statista. 2023. Number of social media users worldwide from 2017 to 2027. *Statista*. **https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users**

Suler J. 2004. The online disinhibition effect. *Cyberpsychol. Behav.* 7(3):321–26

Svolik M. 2020. When polarization trumps civic virtue: partisan conflict and the subversion of democracy by incumbents. *Q. J. Political Sci.* 15:3–31

Tetlock PE, Kristel OV, Elson SB, Green MC, Lerner JS. 2000. The psychology of the unthinkable: taboo trade-offs, forbidden base rates, and heretical counterfactuals. *J. Pers. Soc. Psychol.* 78(5):853–70

Toner K, Leary MR, Asher MW, Jongman-Sereno KP. 2013. Feeling superior is a bipartisan issue: Extremity (not direction) of political views predicts perceived belief superiority. *Psychol. Sci.* 24(12):2454–62

Törnberg P. 2022. How digital media drive affective polarization through partisan sorting. *PNAS* 119(42):e2207159119

Tosi J, Warmke B. 2016. Moral grandstanding. *Philos. Public Affairs* 44(3):197–217

Tosi J, Warmke B. 2020. *Grandstanding: The Use and Abuse of Moral Talk*. Oxford, UK: Oxford Univ. Press

Tucker JA, Theocharis Y, Roberts ME, Barberá P. 2017. From liberation to turmoil: social media and democracy. *J. Democr.* 28(4):46–59

Valentino-DeVries J, Eder S. 2022. For Trump's backers in congress "Devil Terms' help rally voters. *New York Times*, Oct. 23. **www.nytimes.com/2022/10/22/us/politics/republican-election-objectors-rhetoric.html**

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. *J. Commun.* 67(5):803–26

Van Bavel JJ, Baicker K, Boggio PS, Capraro V, Cichocka A, et al. 2020. Using social and behavioural science to support COVID-19 pandemic response. *Nat. Hum. Behav.* 4(5):460–71

Van Bavel JJ, Packer DJ. 2021. *The Power of Us: Harnessing Our Shared Identities to Improve Performance, Increase Cooperation, and Promote Social Harmony*. New York: Little, Brown Spark

Van Bavel JJ, Harris EA, Pärnamets P, Rathje S, Doell KC, Tucker JA. 2021a. Political psychology in the digital (mis) information age: a model of news belief and sharing. *Soc. Issues Policy Rev.* 15(1):84–113

Van Bavel JJ, Packer DJ, Haas IJ, Cunningham WA. 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):e48693

Van Bavel JJ, Pretus C, Rathje S, Pärnamets P, Vlasceanu M, Knowles E. 2023. The costs of polarizing a pandemic: antecedents, consequences, and lessons. *Perspect. Psychol. Sci.* In press. **https://doi.org/10.1177/17456916231190395**

Van Bavel JJ, Rathje S, Harris E, Robertson C, Sternisko A. 2021b. How social media shapes polarization. *Trends Cogn. Sci.* 25(11):913–16

Van de Vyver J, Abrams D. 2015. Testing the prosocial effectiveness of the prototypical moral emotions: Elevation increases benevolent behaviors and outrage increases justice behaviors. *J. Exp. Soc. Psychol.* 58:23–33

van der Linden S, Roozenbeek J, Compton J. 2020. Inoculating against fake news about COVID-19. *Front. Psychol.* 11:566790

van der Linden S, Roozenbeek J, Maertens R, Basol M, Kácha O, et al. 2021. How can psychological science help counter the spread of fake news? *Span. J. Psychol.* 24:E25

Van Geel M, Goemans A, Toprak F, Vedder P. 2017. Which personality traits are related to traditional bullying and cyberbullying? A study with the Big Five, Dark Triad and sadism. *Pers. Individ. Differ.* 106:231–35

van Zomeren M, Spears R, Fischer AH, Leach CW. 2004. Put your money where your mouth is! Explaining collective action tendencies through group-based anger and group efficacy. *J. Pers. Soc. Psychol.* 87(5):649–64

Vogels EA, Perrin A, Anderson M. 2020. Most Americans think social media sites censor political viewpoints. *Pew Research Center*, Aug. 19. **https://www.pewresearch.org/internet/2020/08/19/most-americans-think-social-media-sites-censor-political-viewpoints/**

Vosoughi S, Roy D, Aral S. 2018. The spread of true and false news online. *Science* 359(6380):1146–51

Walter AS, Redlawsk DP. 2019. Voters' partisan responses to politicians' immoral behavior. *Political Psychol.* 40(5):1075–97

Westfall J, Van Boven L, Chambers JR, Judd CM. 2015. Perceiving political polarization in the United States: Party identity strength and attitude extremity exacerbate the perceived partisan divide. *Perspect. Psychol. Sci.* 10(2):145–58

Williams JB, Singh L, Mezey N. 2019. # MeToo as catalyst: a glimpse into 21st century activism. *Univ. Chicago Legal Forum* 2019:22

Wood T, Porter E. 2019. The elusive backfire effect: mass attitudes' steadfast factual adherence. *Political Behav.* 41:135–63

Xiao E, Houser D. 2011. Punish in public. *J. Public Econ.* 95(7–8):1006–17

Yarchi M, Baden C, Kligler-Vilenchik N. 2021. Political polarization on the digital sphere: a cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. *Political Commun.* 38(1–2):98–139

Yang Y, Davis T, Hindman M. 2023. Visual information on Facebook. *J. Commun.* 2023:jqac051

Yildirim MM, Nagler J, Bonneau R, Tucker JA. 2021. Short of suspension: how suspension warnings can reduce hate speech on Twitter. *Perspect. Politics* 21(2):651–63

Zaki J, Bolger N, Ochsner K. 2009. Unpacking the informational bases of empathic accuracy. *Emotion* 9(4):478–87

Zhuravskaya E, Petrova M, Enikolopov R. 2020. Political effects of the Internet and social media. *Annu. Rev. Econ.* 12:415–38

Zimmerman AG, Ybarra GJ. 2016. Online aggression: the influences of anonymity and social modeling. *Psychol. Popul. Media Cult.* 5(2):181–93