PNAS nexus

PNAS Nexus, 2024, **3**, pgae193
https://doi.org/10.1093/pnasnexus/pgae193
Advance access publication 11 June 2024
Review

Morality in the anthropocene: The perversion of compassion and punishment in the online world

Claire E. Robertson[a], Azim Shariff[b] and Jay J. Van Bavel[a,c,d,*]

[a]Department of Psychology, New York University, New York, NY 10003, USA
[b]Department of Psychology, University of British Columbia, Vancouver, BC V6T 1Z4, Canada
[c]Department of Neural Science, New York University, New York, NY 10003, USA
[d]Department of Strategy & Management, Norwegian School of Economics, Bergen 5045, Norway
*To whom correspondence should be addressed: Email: jay.vanbavel@nyu.edu
Edited By: Michele Gelfand

Abstract

Although much of human morality evolved in an environment of small group living, almost 6 billion people use the internet in the modern era. We argue that the technological transformation has created an entirely new ecosystem that is often mismatched with our evolved adaptations for social living. We discuss how evolved responses to moral transgressions, such as compassion for victims of transgressions and punishment of transgressors, are disrupted by two main features of the online context. First, the *scale* of the internet exposes us to an unnaturally large quantity of extreme moral content, causing compassion fatigue and increasing public shaming. Second, the physical and psychological *distance* between moral actors online can lead to ineffective collective action and virtue signaling. We discuss practical implications of these mismatches and suggest directions for future research on morality in the internet era.

Keywords: morality, online mismatches, compassion, punishment and shaming

Morality in the anthropocene: the perversion of compassion and punishment in the online world

Just as the atomic bomb changed how nations conduct warfare and the birth control pill changed how people have sex, the internet has changed moral psychology. The human tendency to care about moral issues like fairness, reciprocity, and empathy were evolutionarily adaptive for improved functioning in small, close-knit societies where people directly relied on their close social ties to survive (1–3). Today, the environment people inhabit is undergoing a shift that is arguably larger than that of the agricultural revolution 12,000 years ago. Estimates suggest that over 5 billion people (over 60% of the entire world) use the internet regularly (4). This number is much higher in developed countries, where rates of regular use are as high as 99%, making the experience of the internet nearly universal in some cultures (5). In this article, we explain how the internet disrupts humanity's basic moral instincts. Our review explains how people's evolved moral psychology makes it easy to exploit them with algorithms, endless newsfeeds, and outrageous content.

The shift to the online environment fundamentally changed the social world, and we argue that evolved behaviors that were advantageous in small groups are often poorly suited to navigate the online environment. Evolved responses to moral conflict between group members, like compassion for the victim and punishment for the transgressor, have different outcomes online than they do in small groups. Here, we discuss how the socially functional outcomes of compassion and punishment are disrupted online by two main features of the online context. First, the *scale* of the internet exposes us to an unnaturally large quantity of extreme moral content. Online, people are exposed to moral content in greater quantities and of greater intensity than they are offline, causing dysfunctional outcomes like compassion fatigue and increasing public shaming. Second, the physical and psychological *distance* between moral actors online makes people's reactions to moral transgressions evolutionarily mismatched. The increased distance between punishers and transgressors online shifts the dynamics of punishment from their evolutionary optima, leading to ineffective collective action and virtue signaling. These mismatches play a role in increasing negativity, outrage, and intergroup conflict (Fig. 1).

Evolutionary underpinnings of moral cognition

Humans are a highly social species (6), and much of the evolved, innate behaviors that humans possess are related to navigating social situations (3, 7–10). People are far more likely to both survive and thrive when they have strong social connections (11). Thus, morality is hypothesized to have evolved due to early humans' need to effectively cooperate with fellow group members

Competing Interest: The authors declare no competing interest.
Received: February 16, 2024. **Accepted:** May 3, 2024

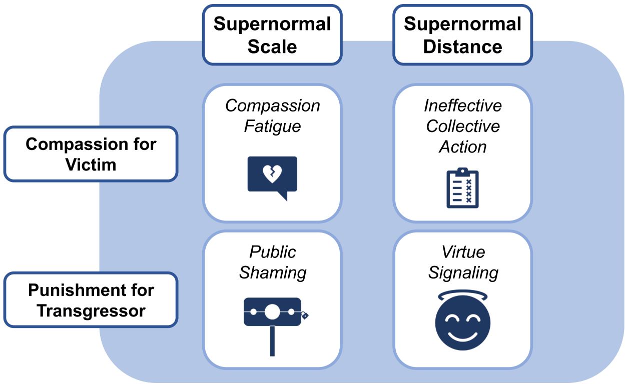

Fig. 1. Visual representation of the framework for how the scale and distance afforded by the internet distorts our evolved reactions for compassion for victims and punishment of transgressors in moral interactions. *Top left*: When the supernormal scale of the internet interacts with people's instinct to feel compassion for victims of moral transgressions, it can result in compassion fatigue. *Top right*: when the supernormal distance of the internet interacts with people's instinct to feel compassion for victims of moral transgressions, it can result in ineffective collective action. *Bottom left*: When the supernormal scale of the internet interacts with people's instinct to punish moral transgressors, it can result in public shaming. *Bottom right*: When the supernormal distance of the internet interacts with people's instinct to punish moral transgressors, it can result in virtue signaling.

and navigate social relationships (3, 12). Violations of cooperative relationships—be it through causing harm, failing to reciprocate, or betraying obligations to a family or group—are seen as morally transgressive (3). The quick recognition of and reaction to moral stimuli is functional, especially in the context of evolutionary adaptation of humans' ancestors (13). In small group contexts, communities of individuals who are predisposed to detect and react negatively to violations of care and cooperation norms are likely to build stronger and more successful groups over time (14–16). A tendency to avoid causing suffering to others and to punish those who cause others suffering bestowed fitness benefits by increasing reciprocity, reducing in-group violence, and signaling positive parental traits. Thus, preferentially attending to moral stimuli elicited helpful and protective behavior, and continues to this day (9, 17–20).

As society became more complex, so too did people's conceptualization of and reasoning about morality. Today, moral reasoning depends on culturally specific norms (21, 22), and occurs via complex cognitive systems by which people blend emotionality and rationality, take context and intentionality into account, and make utilitarian judgments when necessary (23–25). Moreover, it is regulated and guided by institutions and elected third parties (26). Nonetheless, vestiges of people's evolved instincts remain and continue to influence moral cognition and decision making (27–29). For instance, attention towards morally relevant stimuli is hard to suppress—as people recognize morally relevant stimuli more quickly and more consistently than other types of stimuli (30, 31). Other research suggests that moral and emotional language capture early visual attention better than neutral content (32). Thus, people seem to have an attentional preference for content that signals moral relevance.

The internet and supernormal moral stimuli

The modern era of the anthropocene—the epoch of time in which humans have been the dominant force in the global environment (33, 34)—has been likewise marked by a substantial change in the size and complexity of human social networks (35). For almost 99% of our species' history, humans lived in small, nomadic tribes—a state that characterized what is commonly referred to as our Environment of Evolutionary Adaptedness (36). With the Pleistocene–Holocene transition roughly 12,000, humans began to shift away from this state—moving to settled agricultural communities, to market-based economies, and eventually into a communication age driven by recent technologies such as newspapers, telephones, and televised mass media. But the shift to the internet in the last 30 years has fundamentally changed the scale of social interactions and information (37). Unlike the post and telephones which connect people one-to-one, or newspapers and mass media which connect people one-to-many, the internet is the first technology that allows for connections of the many to the many with no concern for time or distance. It has fundamentally changed the way people all over the world communicate with one another. Moreover, it has introduced an entirely new environment—one not just dominated, but wholly created, by human beings.

The internet now connects over 5.3 billion people around the world (38). People spend an average of almost 7 hours per day online, almost as much as the time spent sleeping (39). In those 7 hours, people consume a massive amount of content: data from Facebook suggest that people scroll through roughly 300 feet of content a day, or almost the height of the Statue of Liberty (40). This amount of content is equivalent to reading every page of *The New York Times* more than three times over. It is also orders of magnitude larger than the single newssheets that represent the first iterations of newspapers in the United States in the early 18th century (41). This content comes from many people across distributed social networks that are much larger than previous estimates of historical social network size (8).

Much of the activity that people engage in online relates to social goals (42). As people are exposed to more social content in general, the rate of moral content people are exposed to is also increasing. For instance, people are significantly more likely to learn secondhand about an immoral event in an online context than from print, radio, and TV combined (43) (Fig. 2). This is a striking difference from the infrequency of morality in everyday conversations (44) and underscores the centrality of morality online. We describe two factors that exploit people's attention towards morality in the online environment: overabundance and extremity.

Overabundance

The overabundance of moral content online is likely related to people's attentional preference towards morally relevant stimuli (13). In the attention economy, moral content often generates the greatest engagement (45). For example, tweets that contain moral–emotional language have a greater likelihood of being shared than neutral tweets (46)—this is true for tweets by both lay people and political elites (47). Similarly, news stories that are framed morally receive more shares than neutral news stories online (48). Moreover, the same moral and emotional words that capture attention in controlled lab settings are also more likely to be shared (i.e. retweeted) within real social media contexts (49). Consequently, these results suggest that the attention-grabbing nature of moral and emotional words contributes to the accelerated spread of moral content on social media platforms.

Overabundance of stimuli across many domains can have cognitive consequences due to the way humans detect and

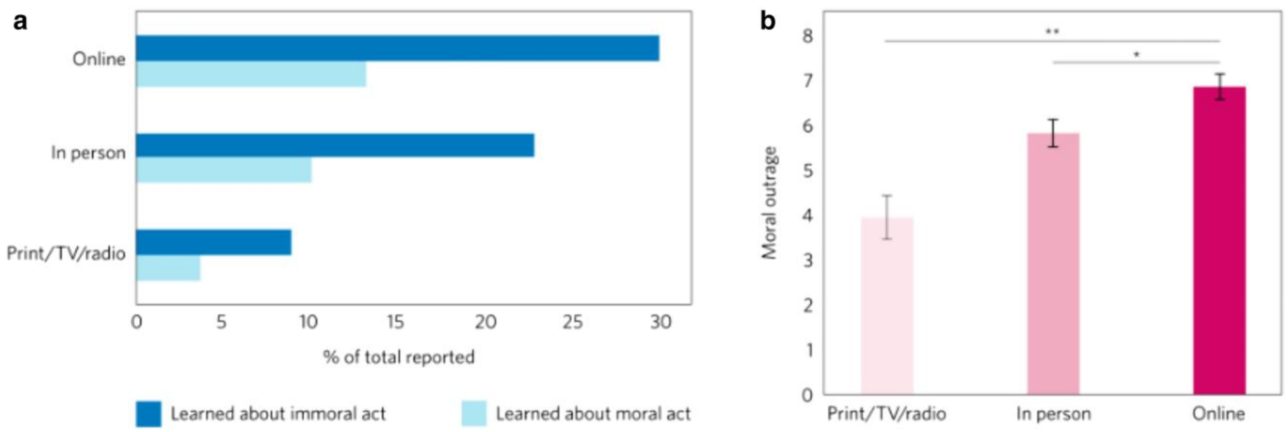

Fig. 2. In a large sample of North American adults, a) People were more likely to learn about immoral acts online than in person or via traditional forms of media (print, television, and radio). The figure displays the percentage of total reported moral/immoral acts that were learned about in each setting. b) Immoral acts encountered online evoked more outrage than immoral acts encountered in person or via traditional forms of media. Error bars represent SEM (Figure adapted from Ref. (43)).

summarize information about others. When a target stimulus is presented rarely, people tend to miss the actual appearance of a stimuli. However, when the target is presented with over-abundance, people tend to report the target even when it is not there (50). In an online environment saturated with moral transgressions, this could lead people to perceive transgressions even when there are none present. Moreover, moral content is prioritized in visual attention, and this predicts online engagement (32). As such, overexposure to moral content might shape behavior in numerous ways.

Regarding information summation, people weigh negative information more heavily than positive information about a person (51). As people summarize information about another person's moral character, negative moral information has a stronger effect on perception of character than positive information (52). Additionally, when people are given many unique exemplars to remember, they engage in a process called ensemble coding, by which they take the average of a series of stimuli based on certain traits (53, 54). However, ensemble coding can be biased by the most extreme or unexpected exemplars in a group (55, 56)—a particular problem online and on social media platforms, where people with the most extreme views generate the most content (57). Indeed, 97% of political posts from Twitter/X come from just 10% of users, meaning that roughly 90% of the population's political opinions are being represented by less than 3% of posts online (58). This reveals how information summation may be misled by the overabundance of information online, leading to biases towards negative moral evaluations or the generation of extreme false norms.

Extremity

The moral content people are exposed to online is often more extreme than typical moral content. The immoral acts that people learn about online tend to elicit stronger feelings of outrage compared to the events that are witnessed in person (43). This suggests that the immoral acts learned online tend to be more extreme than immoral acts encountered in person. One way to think about the effects of heightened extremity of moralized content online is through the lens of *supernormal stimuli*. Supernormal stimuli mimic the stimuli in the environment that organisms are predisposed to preferentially attend to, but are more extreme than they would ever be in the natural environment (42, 59–62). For example, modern fast food is considered to be a supernormal stimulus (63). People evolved to seek out fatty and calorically dense foods, as those types of foods were more likely to help sustain people through periods of relative scarcity that were prevalent in humans' evolutionary history. However, in the modern era of the anthropocene, most people live in relative abundance, and people's tendency towards fatty foods now contributes to people overeating unhealthy foods, leading to heart disease, diabetes, and other health complications. Extreme moralized content online may function in a similar way, capturing our attention and triggering unhealthy behavior against our better judgment.

Recently, Bor and Peterson (64) argued that the mismatch hypothesis does not explain online hostility. They note that people are consistent in their levels of hostility both online and offline, suggesting that online contexts do not change people's hostility, but simply enhance the visibility of people who are already hostile. We argue, however, that the increase in visibility makes the online environment a more routinely hostile and extreme place, potentially creating a mismatch with people's experiences in the real world where such hostility is less visible. Most social media content is produced by a small subset of users who tend to be the most ideologically extreme and the most active online (58,65). Indeed, those who have the strongest negative feelings about a group or topic are also the most likely to share negative content online (66). This may lead the online environment to be saturated by the extreme content posted by those who, in turn, hold the most extreme opinions.

This feature of the online world can artificially inflate people's perceptions of animosity and outrage—creating false norms (49). This may be further distorted because people engage in both homophily where they choose to connect with individuals who are ideologically similar to them (46, 67)—and acrophily where they choose those who share their ideology but are slightly more extreme than them (68). Thus, people's social networks tend to be flooded with opinions that are, on average, more extreme than their own opinions or the opinions they experience in the real world. This is further exacerbated by both algorithms and social reinforcement learning (49, 69). This is a cyclical process: algorithms are built to maximize engagement online, and the people who engage the most are also those with the most extreme opinions (70). Thus, algorithms "learn" that the most extreme content is the most successful at garnering online engagement, and prioritize that type of content—even if people do not like it (see Ref. (71)). We argue that two of the outcomes of this cyclical increase in extreme content are the disruption of compassion and third-party punishment online.

In summary, we theorize that the overabundance and extremity of online content lead people's evolutionary moral dispositions to be perpetually triggered. This, in turn, increases the production and spread of moral content online—further feeding a morally saturated environment. In the next two sections, we examine two areas of moral cognition—compassion and third-party punishment—to illuminate how the internet and social media exploits basic moral cognition, eliciting behavior that is maladaptive for both individuals and society.

Compassion and empathy

Offline and online

It is natural to feel compassion and empathy for victims. In reaction to witnessing a moral transgression, people feel compassion, empathy, and a desire for restitution for the victim (9, 72, 73). Empathy spurs action—groups whose members can empathize and have compassion for others are more likely to take care of each other and of vulnerable offspring, increasing the odds of survival and gene propagation (9, 74). In modern times, empathy is associated with higher donations to charity and those in need (75, 76). However, the compassion that humans evolved to feel for victims is altered due to the distance between social ties online.

Despite these benefits, people are selective in whom they empathize with (77). People are more likely to empathize with in-group members compared to out-group members (78, 79) and less likely to feel empathy for more distant social connections (80). This is because empathizing can be emotionally taxing, and people will avoid it when possible (81). Moreover, empathy is a costly cognitive resource, and people want to reserve it for those who may be able to help them at a later time, such as in-group members (20). Thus, the limits of empathy are regularly tested in online contexts, where people are exposed to supernormal levels of moral content from distant and loose social connections.

When people are overloaded with requests for empathy, people find assigning blame easier than having empathy (82). Online, this may lead to people reacting to transgressions to focus on assigning blame rather than empathizing with a victim. This is especially problematic, since one of the most effective ways to reduce hateful speech online is to express empathy (83). When comparing online empathy and offline empathy directly, offline empathy is significantly stronger than online empathy (84), suggesting that people may morally disengage online, relieving themselves of the responsibility to act (85). Taken together, this evidence suggests that people are less likely to feel compassion and act in restorative ways, and more likely to assign blame to victims when confronted with the supernormal quantities of suffering that are typical of online engagement.

Supernormal scale and compassion fatigue

The tendency to feel compassion towards the victims of a moral transgression does not scale well online due to the high exposure to victims. People respond with more empathy to a single victim than to a group of victims (86, 87). They become numb to excess suffering and do not linearly scale their empathy with the number of victims. For example, people are willing to donate roughly the same amount of money to help from 2,000 to 200,000 victims (88). This may be in part because people are averse to taking on too much responsibility for large numbers of moral victims (89). Indeed, when there are many victims rather than just a few, people are motivated to disengage from a conflict and not act (90). As the number of victims in a scenario increases, the likelihood that people will take prosocial action like donating money actually goes down (90). This may be related to processes by which, when an experience is common, people value it less over time than when it is rare (91). For example, overexposure to moral transgressions can have a numbing effect on observers. When people are repeatedly exposed to the same information about a moral transgression, they later report that that transgression seems less unethical than a novel transgression (92). This may lead them to feeling that the transgression was "not that bad" and therefore reduce their compassion for a victim.

Even when people do choose to behave prosocially online, their actions often make little to no real impact. This may be because of moral licensing, or the belief that a prior good deed "licenses" a person to engage in morally questionable behavior later (93). For example, engaging in a noncostly form of compassion, such as "liking" or "sharing" a post, may lead people to believe that they have absolved themselves of their moral responsibility to engage in further prosocial action (94, 95). Indeed, the common tagline of "thoughts and prayers," often posted online after disasters in the United States, may undercut monetary donations to those in need (95). There are exceptions to this—the Ice Bucket Challenge, for example, raised millions of dollars for ALS research, and relied on people's desires to share prosocial information online (96). In most other cases, however, low-cost forms of prosocial behavior can, ironically, hinder the material impacts of positive social movements. Thus, an evolved tendency for compassion and empathy can lead to a decrease in overall prosocial behavior when in an online context.

Supernormal distance and ineffective collective action

In rare cases, mass sharing can be helpful. Internet use has been credited with spawning protests and demonstrations of collective action around the globe such as the Arab Spring and Black Lives Matter (97, 98). The internet has indisputably increased broad awareness of a wide variety of social issues. The virality and traction these issues received, especially as they gave suppressed voices who may have been typically ignored by mainstream media an outlet to collectively organize and share experiences (99), created broad awareness that would have been impossible without the internet. Unfortunately, while awareness of social issues is often a net positive, it does not directly translate to increased action towards fixing an issue. Indeed, there is increasing debate about how (in)effective online activism really is (100, 101). For instance, even though social media-driven nonviolent protests are larger now compared to most historical protest movements, they have resulted in far less policy change (102). This may be in part because of increased psychological distance between individuals who participate in online activism (101, 102). This has led to broad but shallow interest in these causes which may actually harm the causes in the long run (103) and foster cynicism.

It is theorized that this drop in efficacy is because activism used to require deep roots and structures to get off the ground, stronger dedication to a cause, and months of planning to execute (101). This led to vibrant social networks and clear organizational goals. Now protests can be organized within a matter of days due to social media, potentially leading more people to show up to a protest (98). However, many of those who attend protests are less dedicated to the cause or the group than they would have been the case historically. Moreover, their engagement might be motivated by superficial self-interest (e.g. creating online content to signal an affiliation to gain social status). Thus, while online activism may increase awareness of inequities or social problems, it can actually hinder the effectiveness of collective action by prioritizing shallow, low-cost forms of collective action that are not effective at convincing or pressuring those in power to make lasting policy change (101, 102, 104).

Third-party punishment

Offline and online

In addition to compassion for the victim, witnessing a moral transgression also spurs punishment towards the transgressor. Like compassion fatigue, the desire to punish a wrongdoer often occurs when one is a third party to a moral transgression. In fact, people are most punitive when they are mere bystanders to a moral transgression (105). The drive to engage in costly third-party punishment—or the act of punishing wrongdoers even when that punishment comes at a personal cost—appears to be a culturally universal, likely evolved, tendency (7, 18, 106). Research using economic games has found robust evidence of third-party punishment (107). The motivation to punish transgressors emerges early in development, as young children engage in costly punishment towards in-group and out-group moral transgressors (108) giving up a treasured resource (being able to use a slide) in order to punish other children who behaved immorally (109). Evolutionarily, costly third-party punishment may have developed in small groups to deter cheating and freeriding behavior, thus strengthening the group over time (7, 14, 15).

On the surface, third-party punishment is an evolutionary puzzle: why would it be beneficial to sacrifice one's own resources to punish a bad actor, especially when one is not personally harmed? One clue is that third-party punishment is only effective when people are required to cooperate with the same group repeatedly (14). Punishment is not as effective when the makeup of groups changes, or in one-shot dilemmas. Furthermore, punishment increases cooperation and group resource contributions most when it is done in public, or in full view of the rest of the group (110). This suggests that social shame acts as a deterrent for bad behavior among in-group members, in addition to any material loss incurred as punishment. Additionally, publicly rebuffing someone helps the punisher by deterring future cheaters (110). Thus, punishment as a response to witnessing moral transgressions highlights deep-rooted motivations to punish wrongdoers and uphold fairness in social interactions.

In addition to punishing cheaters to deter future immoral behavior, engaging in third-party punishment may confer reputational benefits to the punisher (111, 112). Indeed, people are more likely to engage in third-party punishment when they have an audience (19). Part of the reason that third-party punishment is effective at maintaining group cohesion is that it signals commitment to one's group and re-establishes that commitment as a group norm. To be effective, it requires a real sacrifice, either in resources or in personal risk, to the group for the sake of justice (15, 18, 110, 113–115). Thus, engaging in third-party punishment makes someone a more attractive mate or cooperation partner, as it signals trustworthiness and willingness to sacrifice for others (16). Indeed, engaging in costly third-party punishment demonstrates moral fiber to one's group members, and can lead to admiration and increased status in the eyes of observers (110–112). Computer models of evolving group dynamics found that group members who remained in "good standing" reputationally (i.e. helped others when they could) propagated their genes more easily over time (10).

Supernormal scale and public shaming

When this tendency to punish moral wrongdoers is engaged in the online context, it has unexpected consequences. As the number of possible third-party punishers increases, the average third-party punishment intensity decreases only mildly, leading to a substantial increase in total punishment as group size increases (116). When people learn of a moral transgression online, they have an urge to punish the transgressor, just as was the case when punishment occurred in small groups. However, online interactions do not take place within a small group. On the contrary, many instances of online shaming or punishment involve one transgressor being punished by thousands of people, most of whom have no offline relationship with the transgressor (117). People in online communities are not required to work or live together at any point, because they are geographically spread apart and do not visibly rely on one another to fulfill day to day tasks. Online, groups function more to signal belonging to a specific social identity such as political party. The superficiality of these connections to relatively unknown strangers can lead people to have black-and-white judgments of morality with little nuance (118). This can lead to a massive campaign of retribution against a complete stranger.

Due to the massive scale of online social networks, the population from which third-party punishment can spring is immense. Instead of a small tribe of people who have a vested interest in fostering group cooperation, millions of people from anywhere in the world can gather to publicly punish one person with no personal investment or genuine desire for restitution. They might seek to gain social status without any genuine attempt to improve collective outcomes. Throughout evolutionary history, third-party punishment was usually administered by people who had a stake in the outcome, and also typically by in-group members. Indeed, the likelihood that a third-party observer would eventually have to interact with either the transgressor or the victim of a moral transgression was extremely high (119, 120). However, in online contexts there are millions of third-party observers, and very few, if any, will ever meet a particular transgressor in real life. This can undercut the traditional social function of cooperation and incentivize activities like public shaming that are disproportionate to the original transgression. Punishment in this context focuses on exacting retribution instead of rehabilitation or education.

Supernormal distance and virtue signaling

Physical distance between the punisher and the punished means that online shaming and punishment is rarely costly to the punisher (43). Thus, punishing people online is not an effective signal of group commitment or trustworthiness. As such, third-party punishers may engage immoral grandstanding or selfish virtue signaling (121, 122). In the online environment, virtue signaling refers to a type of false signaling where people publicly claim to be morally virtuous to enhance their own moral reputation, without exemplifying that virtue in a meaningful way (123). Online, there is near endless evidence of out-group members behaving badly, allowing in-group members many opportunities to signal their status as "good group member" and respond virtuously, inadvertently escalating the conflict (124). This can undercut the core function of costly punishment by making it cheap enough for noninvested strangers to participate.

Importantly, people can signal their true moral beliefs on social media. However, virtue signaling is often seen as hypocritical in online contexts because the signaler received social rewards (i.e. likes/shares) for saying the "right thing" without requiring the signaler to actually "do the right thing" (125). Thus, when moral outrage and shaming goes viral, and thousands of people costlessly reprimand a single transgressor, outside observers are less likely to see that outrage as genuine (126). Instead, people perceive punishers as bullies when they are part of a large group of online punishers and begin feeling empathy for the original transgressor. Hence, people's evolutionary motives to punish moral transgressors may have an inverse effect from their evolved function: rather than signaling that one is just and righteous, others may perceive their virtue signaling as a sign of immorality (126) or disingenuousness (127). Thus, online public shaming can have the opposite effect from its evolutionary roots, reducing trust in punishers and increasing sympathy for transgressors. It may also foster genuine cynicism about the actors or about online moral discourse.

Regarding the shaming and punishing of a moral transgressor, the evolutionary mismatch of punishment tendencies in the new online context changes the outcome of punishment. In addition to increasing the status of a punisher, third-party punishment also served evolutionarily to deter cheaters from transgressing again (128, 129). However, the deterring effects of punishment and shaming worked best when engaged in small groups who would have repeated interactions over time (14, 130). The online context is different in both of these regards. Due to the extremely high rate of relational mobility (i.e. the frequency and flexibility by which people are able to encounter new social partners, and form and end social relationships) that people experience online, they are easily able to move out of one group and into a new group with all new social participants (131–133). As a result, punishing transgressors may not successfully deter repeated wrongdoings when executed online. Therefore, people feel more at liberty to say or do things online that they would not say in real life (134, 135).

Furthermore, public shaming transgressors may actually increase their negative feelings and resentment towards punishers, rather than guilt over their transgressive actions (136, 137). This may lead transgressors to focus on the proportionality of their transgression compared to the reaction of the public, rather than on changing their behavior (117). This can lead to the continuation or escalation of conflict. Transgressors might even develop communities around these grievances and seek revenge. Thus, the shifting dynamics of the online realm, characterized by high relational mobility and the perception of punishers as bullies, reduce the effectiveness of punishment as a deterrent against repeated wrongdoings.

Future directions

We have presented several clear examples where we think mismatches lead to surprising patterns of behavior. Research is now needed to test whether the assumptions made by the mismatch hypothesis are supported by empirical evidence. For example, if people engage in public shaming in order to reap reputational benefits from engaging in costly punishment (16), do people also go out of their way to signal that their online punishment was somehow costly to them despite the distance between themselves and the punished? Furthermore, if people engage in ineffective prosociality online to morally license themselves to disengage from mass suffering (89, 95), do people feel less empathy for victims of moral transgressions after they have been given a costless opportunity to express compassion on social media? Relatedly, given the ephemeral nature of online activism (101, 103), did movements that called for a long-term offline commitment to a cause, such as a boycott of a product or store, result in greater behavioral and psychological commitment to the cause compared to causes focusing on shorter offline commitments, such as protests? With these insights, researchers can begin to develop interventions to reduce negative individual and societal outcomes related to compassion and punishment mismatches online.

Part of the problem with reducing the mismatch between evolved moral behavior and the online environment is that the attention economy upon which the internet is built is currently structured to incentivise supernormal stimuli (45, 138, 139). The online environment is owned and regulated by a number of technology companies whose primary profits come from advertising (140). Advertising requires that people are engaging on a specific platform, and tech companies must compete for user attention (140). Considering that moral content often receives preferential attention (32), it is logical for companies to capitalize and promote moral content. There is little financial incentive for companies who profit from attention capture to reduce the use of supernormal stimuli on their platforms. For instance, interventions that reduce one's exposure to toxicity online also reduce engagement on social media sites (141). This undercuts the profitability of these platforms. Thus, it is unlikely companies will be motivated to change the online context in ways that ameliorate these evolutionary mismatches (without government regulation). On the contrary, we think it is more likely that companies will continue to exploit these tendencies as long as it remains profitable.

As such, future research should test platform design features that sustain attention or engagement without inducing negative externalities on individuals and society. There is evidence that people have a desire to make the internet a more positive place but lack the means to do so on their own. When asked directly, most online media users say that they want lower levels of outrage and negativity in their online feeds (71). Thus, allowing people to more easily regulate the types of content they do and do not want to see may reduce people's baseline exposure to morally outrageous content (139). Other design changes, such as allowing people to publicly signal their "trust" of a particular news story as an alternative to "liking" or "sharing" news, may help reduce the spread of misinformation by downregulating attention-grabbing and morally stimulating headlines (138). More research is needed on these prosocial design features.

Future research should focus on the longitudinal effects of overexposure to moral information online, especially looking at individual differences. Prior work examining individual-level outcomes in overexposure to the internet and social media have found that, while social media use can be positive for some people, it can have negative effects for vulnerable or at-risk populations (142). Moral discourse online is linked to subsequent violence in the real world (143). Furthermore, the internet has been a boon for hate groups–allowing them to flourish and organize extremists (144). Critically, even though certain conspiracy theories may originate online, they often bleed into the offline world, causing extremism, harm, and even death (145, 146).

While these studies have been correlational, large-scale experiments have found that limiting social media causes improvements in subjective well-being (e.g. (147, 148)). Thus, researchers should examine whether full social media cessation is required for well-being improvements, or if removal or reduction of specific content such as extreme content or users, could allow people to continue using social media while still improving their social interactions and well-being. Much is still unknown about the long-term effects of overexposure to negative information online.

It is difficult for researchers to effectively study the online environment because tech companies are reluctant to share how their algorithms function (149). This is true even though there is widespread agreement among lay people that greater transparency about social media algorithms (71). This critical lack of understanding has hindered scientists' abilities to critically examine the effects of social media on emotion and behavior (140, 150). Therefore, it is imperative that researchers have greater access to these algorithms to develop a better understanding of how they function. Ideally, stakeholders (e.g. users and members of the public) should also have input into algorithms that impact their lives.

We acknowledge that both the effect of social media and evolutionary theory are hard to test experimentally. One cannot assign people to have zero exposure to social media, or to acquire a specific evolutionary adaptation. Instead, much research on the effects of social media are correlational, or rely on natural experiments from archival data (for an example, see (151)). In order to drill down on the causal and evolutionary mechanisms that contribute to the mismatch of moral instincts online, researchers should consider more ambitious methods. Causal social media studies, such as cessation studies, have been effective in the past at investigating social media's effect on polarization in the United States (147) and Bosnia and Herzegovina (148). However, in order to argue that a trait is evolved and not learned, there must be evidence of that trait across cultures. Global collaborative efforts to replicate these studies and examine a wider range of outcome measures, including moral outrage and extremism, are already underway.[a]

Although we have focused on the areas where the distance between the traditional offline environment and the new online one has undermined the effectiveness of compassion and punishment, the internet is obviously not all bad. For example, the scale of the internet has raised the ceiling and lowered barriers for nearly every type of human knowledge, from simple online tutorials for learning new skills (152) to crowdsourcing solutions to our most difficult and pressing scientific conundrums (153). Furthermore, although people are more distant from those in their social groups, the internet has also brought together new social groups that could never have existed before, such as support groups for people with rare diseases who would have been unlikely to find each other in the real world due to physical distance (154, 155). While these benefits may be clear and demonstrable, however, the internet has also led to unexpected but consistent consequences that must be investigated as well. While small support groups may be positive forces in the lives of their users, why do large-scale social movements that originate online often stagnate (101, 102)? Why, when high quality knowledge is now universally available, does fake news proliferate online (156, 157)? Why are social media users willing to pay to have others—including themselves—deactivate these popular social media platforms (i.e. TikTok and Instagram; (158))? Understanding how the structure of the online environment can lead to such negative outcomes is the crucial first step in developing interventions and solutions to mitigate those negative outcomes.

Conclusion

The changes that the internet has caused to our social environment have been larger and faster than any cultural or technological shift in our history. Humans are left using brains tuned for an offline world to navigate a novel environment of extreme stimuli and connectedness. However, humans have also evolved to be keen social learners and remarkably adaptable (159). Understanding how the internet can distort our moral instincts will help us navigate and shape our new environment and help prevent maladaptive outcomes for individuals and society.

Notes

[a] https://globalsocialmediastudy.com/

Funding

This research was supported by a Canada 150 Research Chair grant from the Social Science and Humanities Research Council of Canada to A.S., and a Google Jigsaw grant, a Russell Sage Foundation grant (G-2110-33990), and a grant from the John Templeton World Charity Foundation awarded to J.V.B. (TWCF-2022-30561).

Author Contributions

C.R.: conceptualization; visualization; writing-original draft; writingreview and editing. A.S.: conceptualization; writing-review and editing. J.V.B.: conceptualization; writing-review and editing.

Preprints

This manuscript was posted as a preprint: doi:10.31234/osf.io/ns34y.

References

1 Darwin C. 1871. *The descent of man and evolution in relation to sex*. London: Murray.

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

3 Krebs DL. 2008. Morality: an evolutionary account. *Perspect Psychol Sci*. 3(3):149–172.

4 Petrosyan A. 2024. Internet and social media users in the world 2022. Statista [accessed 2022 Sep 30]. https://www.statista.com/statistics/617136/digital-population-worldwide/.

5 Oberlo. 2022. Internet use by country [updated Aug 2022] Oberlo [accessed 2022 Sep 27]. https://www.oberlo.com/statistics/internet-use-by-country.

6 Tomasello M. 2014. The ultra-social animal. *Eur J Soc Psychol*. 44(3):187–194.

7 Bowles S, Gintis H. 2004. The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theor Popul Biol*. 65(1):17–28.

8 Dunbar RIM. 1993. Coevolution of neocortical size, group size and language in humans. *Behav Brain Sci*. 16(4):681–694.

9 Goetz JL, Keltner D, Simon-Thomas E. 2010. Compassion: an evolutionary analysis and empirical review. *Psychol Bull*. 136(3):351–374.

10 Leimar O, Hammerstein P. 2001. Evolution of cooperation through indirect reciprocity. *Proc Roy Soc London, Ser B, Biol Sci*. 268(1468):745–753.

11 Holt-Lunstad J, Smith TB, Layton JB. 2010. Social relationships and mortality risk: a meta-analytic review. *PLoS Med*. 7(7): e1000316.

12 Curry OS, Mullins DA, Whitehouse H. 2019. Is it good to cooperate?: Testing the theory of morality-as-cooperation in 60 societies. *Curr Anthropol*. 60(1):47–69.

13 Gantman AP, Van Bavel JJ. 2015. Moral perception. *Trends Cogn Sci (Regul Ed)*. 19(11):631–633.

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

15 Boyd R, Gintis H, Bowles S, Richerson PJ. 2003. The evolution of altruistic punishment. *Proc Natl Acad Sci USA*. 100(6):3531–3535.

16 Jordan JJ, Hoffman M, Bloom P, Rand DG. 2016. Third-party punishment as a costly signal of trustworthiness. *Nature*. 530(7591): 473–476.

17 Haidt J, Koller SH, Dias MG. 1993. Affect, culture, and morality, or is it wrong to eat your dog? *J Pers Soc Psychol*. 65(4):613–628.

18 Henrich J, *et al*. 2006. Costly punishment across human societies. *Science*. 312(5781):1767–1770.

19 Kurzban R, Descioli P, Obrien E. 2007. Audience effects on moralistic punishment★. *Evol Hum Behav*. 28(2):75–84.

20 Zaki J. 2014. Empathy: a motivated account. *Psychol Bull*. 140(6): 1608–1647.

21 Awad E, Dsouza S, Shariff A, Rahwan I, Bonnefon J-F. 2020. Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proc Natl Acad Sci USA*. 117(5): 2332–2337.

22 Enke B. 2019. Kinship, cooperation, and the evolution of moral systems*. *Q J Econ*. 134(2):953–1019.

23 Gray K, Young L, Waytz A. 2012. Mind perception is the essence of morality. *Psychol Inq*. 23(2):101–124.

24 Greene JD, Nystrom LE, Engell AD, Darley JM, Cohen JD. 2004. The neural bases of cognitive conflict and control in moral judgment. *Neuron*. 44(2):389–400.

25 Schein C, Gray K. 2018. The theory of dyadic morality: reinventing moral judgment by redefining harm. *Pers Soc Psychol Rev*. 22(1):32–70.

26 Van Bavel JJ, Pärnamets P, Reinero DA, Packer D. 2022. Chapter Two—How neurons, norms, and institutions shape group cooperation. In: Gawronski B, editor. *Advances in experimental social psychology. vol. 66*. New York (NY): Academic Press. p. 59–105.

27 Greene J, Haidt J. 2002. How (and where) does moral judgment work? *Trends Cogn Sci (Regul Ed)*. 6(12):517–523.

28 Moll J, Zahn R, de Oliveira-Souza R, Krueger F, Grafman J. 2005. The neural basis of human moral cognition. *Nat Rev Neurosci*. 6(10):799–809.

29 Van Bavel JJ, FeldmanHall O, Mende-Siedlecki P. 2015. The neuroscience of moral cognition: from dual processes to dynamic systems. *Curr Opin Psychol*. 6:167–172.

30 De Freitas J, Hafri A. 2024. Moral thin-slicing: forming moral impressions from a brief glance. *J Exp Soc Psychol*. 112:104588.

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

32 Brady WJ, Gantman AP, Van Bavel JJ. 2020. Attentional capture helps explain why moral and emotional content go viral. *J Exp Psychol Gen*. 149(4):746–756.

33 Lewis SL, Maslin MA. 2015. Defining the anthropocene. *Nature*. 519(7542):171–180.

34 Waters CN, *et al*. 2016. The anthropocene is functionally and stratigraphically distinct from the holocene. *Science*. 351(6269): aad2622.

35 Dunbar RI. 2016. The social brain hypothesis and human evolution. In Oxford research encyclopedia of psychology. https://doi.org/10.1093/acrefore/9780190236557.013.44.

36 Foley R. 1995. The adaptive legacy of human evolution: a search for the environment of evolutionary adaptedness. *Evol Anthropol*. 4(6):194–203.

37 Firth J, *et al*. 2019. The “online brain”: how the internet may be changing our cognition. *World Psychiatry*. 18(2):119–129.

38 Degenhard J. 2024. Global: internet users 2013–2028. Statista [accessed 2023 Jul 24]. https://www.statista.com/forecasts/1146844/internet-users-in-the-world.

39 Kemp S. 2022. Digital 2022: global overview report—DataReportal—global digital insights. DataReportal [accessed 2023 Nov 8]. https://datareportal.com/reports/digital-2022-global-overview-report.

40 Morant L. 2019. The truth behind 6 second ads. Medium [accessed 2023 Nov 8]. https://medium.com/@Lyndon/the-tyranny-of-six-seconds-592b94160877.

41 Park RE. 1923. The natural history of the newspaper. *Am J Sociol*. 29(3):273–289.

42 Tamir DI, Ward AF. 2015. Old desires, new media. In: Hofmann W, Nordgren LF, editors. *The psychology of desire*. New York (NY): Guilford Press. p. 432–455.

43 Crockett MJ. 2017. Moral outrage in the digital age. *Nat Hum Behav*. 1(11):769–771.

44 Atari M, *et al*. 2023. The paucity of morality in everyday talk. *Sci Rep*. 13(1):5967.

45 Van Bavel JJ, Robertson CE, del Rosario K, Rasmussen J, Rathje S. 2024. Social media and morality. *Annu Rev Psychol*. 75(1): 311–340.

46 Brady WJ, Wills JA, Jost JT, Tucker JA, Bavel JJV. 2017. Emotion shapes the diffusion of moralized content in social networks. *Proc Natl Acad Sci USA*. 114(28):7313–7318.

47 Brady WJ, Wills JA, Burkart D, Jost JT, Van Bavel JJ. 2019. An ideological asymmetry in the diffusion of moralized content on social media among political leaders. *J Exp Psychol Gen*. 148(10): 1802–1813.

48 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):803–826.

49 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.

50 Wolfe JM, Van Wert MJ. 2010. Varying target prevalence reveals two dissociable decision criteria in visual search. *Curr Biol*. 20(2): 121–124.

51 Ito TA, Larsen JT, Smith NK, Cacioppo JT. 1998. Negative information weighs more heavily on the brain: the negativity bias in evaluative categorizations. *J Pers Soc Psychol*. 75(4):887–900.

52 Klein N, O’Brien E. 2016. The tipping point of moral change: when do good and bad acts make good and bad actors? *Soc Cogn*. 34(2):149–166.

53 Alvarez GA. 2011. Representing multiple objects as an ensemble enhances visual cognition. *Trends Cogn Sci (Regul Ed)*. 15(3): 122–131.

54 Parkes L, Lund J, Angelucci A, Solomon JA, Morgan M. 2001. Compulsory averaging of crowded orientation signals in human vision. *Nat Neurosci*. 4(7):739–744.

55 Goldenberg A, *et al*. 2022. Amplification in the evaluation of multiple emotional expressions over time. *Nat Hum Behav*. 6(10):1408–1416.

56 Kardosh R, Sklar AY, Goldstein A, Pertzov Y, Hassin RR. 2022. Minority salience and the overestimation of individuals from minority groups in perception and memory. *Proc Natl Acad Sci USA*. 119(12):e2116884119.

57 Robertson C, del Rosario K, Van Bavel JJ. 2024. Inside the funhouse mirror factory: how social media distorts perceptions of norms [accessed 2024 Apr 8]. https://osf.io/kgcrq/download/.

58 Hughes A. 2019. A small group of prolific users account for a majority of political tweets sent by U.S. adults. Pew Research Center [accessed 2023 Nov 16]. https://www.pewresearch.org/short-reads/2019/10/23/a-small-group-of-prolific-users-account-for-a-majority-of-political-tweets-sent-by-u-s-adults/.

59 Barrett D. 2010. *Supernormal stimuli: how primal urges overran their evolutionary purpose*. New York: W. W. Norton & Company, Inc.

60 Tinbergen N. 1948. Social releasers and the experimental method required for their study. *Wilson Bulletin*. 60(1):6–51.

61 Tinbergen N, Perdeck AC. 1951. On the stimulus situation releasing the begging response in the newly hatched herring gull chick (*Larus argentatus argentatus* Pont.). *Behaviour*. 3(1):1–39.

62 Ward AF. 2013. Supernormal: how the internet is changing our memories and our minds. *Psychol Inq*. 24(4):341–348.

63 Barrett D. 2007. *Waistland: the (r)evolutionary science behind our weight and fitness crisis*. 1st ed. New York (NY): W.W. Norton & Co.

64 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.

65 Barberá P, Rivero G. 2015. Understanding the political representativeness of twitter users. *Soc Sci Comput Rev*. 33(6):712–729.

66 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.

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

68 Goldenberg A, *et al*. 2023. Homophily and acrophily as drivers of political segregation. *Nat Hum Behav*. 7:219–230.

69 Brady WJ, Jackson JC, Lindström B, Crockett MJ. 2023. Algorithm-mediated social learning in online social networks. *Trends Cogn Sci (Regul Ed)*. 27(10):947–960.

70 Milli S, Carroll M, Pandey S, Wang Y, Dragan AD. 2023. Engagement, user satisfaction, and the amplification of divisive content on social media. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.16941.

71 Rathje S, Robertson C, Brady WJ, Van Bavel JJ. 2023. People think that social media platforms do (but should not) amplify divisive content. *Perspect Psychol Sci*. 17456916231190392.

72 Batson CD. 1987. Prosocial motivation: is it ever truly altruistic? In: Berkowitz L, editors. *Advances in experimental social Psychology. vol. 20*. San Diego (CA): Elsevier. p. 65–122.

73 Nussbaum M. 1996. Compassion: the basic social emotion. *Soc Philos Policy*. 13(1):27–58.

74 Hoffman ML. 1981. Is altruism part of human nature? *J Pers Soc Psychol*. 40(1):121–137.

75 Fisher RJ, Vandenbosch M, Antia KD. 2008. An empathy-helping perspective on consumers' responses to fund-raising appeals. *J Consum Res*. 35(3):519–531.

76 Zhou X, Wildschut T, Sedikides C, Shi K, Feng C. 2012. Nostalgia: the gift that keeps on giving. *J Consum Res*. 39(1):39–50.

77 Bloom P. 2016. *Against empathy: the case for rational compassion*. 1st ed. New York (NY): Ecco, an imprint of HarperCollins.

78 Cikara M, Bruneau E, Van Bavel JJ, Saxe R. 2014. Their pain gives us pleasure: how intergroup dynamics shape empathic failures and counter-empathic responses. *J Exp Soc Psychol*. 55:110–125.

79 Meyer ML, *et al*. 2013. Empathy for the social suffering of friends and strangers recruits distinct patterns of brain activation. *Soc Cogn Affect Neurosci*. 8(4):446–454.

80 Depow GJ, Francis Z, Inzlicht M. 2021. The experience of empathy in everyday life. *Psychol Sci*. 32(8):1198–1213.

81 Cameron CD, *et al*. 2019. Empathy is hard work: people choose to avoid empathy because of its cognitive costs. *J Exp Psychol Gen*. 148(6):962–976.

82 Bambrah V, Cameron CD, Inzlicht M. 2022. Outrage fatigue? Cognitive costs and decisions to blame. *Motiv Emot*. 46(2):171–196.

83 Hangartner D, *et al*. 2021. Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proc Natl Acad Sci USA*. 118(50):e2116310118.

84 Carrier LM, Spradlin A, Bunce JP, Rosen LD. 2015. Virtual empathy: positive and negative impacts of going online upon empathy in young adults. *Comput Human Behav*. 52:39–48.

85 Bandura A. 2002. Selective moral disengagement in the exercise of moral agency. *J Moral Educ*. 31(2):101–119.

86 Kogut T, Ritov I. 2005. The "identified victim" effect: an identified group, or just a single individual? *J Behav Decis Mak*. 18(3):157–167.

87 Slovic P, Västfjäll D, Erlandsson A, Gregory R. 2017. Iconic photographs and the ebb and flow of empathic response to humanitarian disasters. *Proc Natl Acad Sci USA*. 114(4):640–644.

88 Caviola L, Schubert S, Greene JD. 2021. The psychology of (in)effective altruism. *Trends Cogn Sci (Regul Ed)*. 25(7):596–607.

89 Slovic P. 2007. "If I look at the mass I will never act": psychic numbing and genocide. *Judgm Decis Mak*. 2(2):79–95.

90 Västfjäll D, Slovic P, Mayorga M, Peters E. 2014. Compassion fade: affect and charity are greatest for a single child in need. *PLoS One*. 9(6):e100115.

91 Quoidbach J, Dunn EW. 2013. Give it up: a strategy for combating hedonic adaptation. *Soc Psychol Personal Sci*. 4(5):563–568.

92 Pillai RM, Fazio LK, Effron DA. 2023. Repeatedly encountered descriptions of wrongdoing seem more true but less unethical: evidence in a naturalistic setting. *Psychol Sci*. 34(8):863–874.

93 Blanken I, Van De Ven N, Zeelenberg M. 2015. A meta-analytic review of moral licensing. *Pers Soc Psychol Bull*. 41(4):540–558.

94 Riley CA. 2022. You don't need to post about every tragedy. The Atlantic [accessed 2024 Jan 30]. https://www.theatlantic.com/family/archive/2022/03/social-media-activism-ukraine-solidarity/629402/.

95 Thunström L. 2020. Thoughts and prayers—do they crowd out charity donations? *J Risk Uncertain*. 60(1):1–28.

96 Van Der Linden S, *et al*. 2021. How can psychological science help counter the spread of fake news? *Span J Psychol*. 24:e25.

97 Taylor K-Y. 2016. *From #BlackLivesMatter to black liberation*. Chicago (IL): Haymarket Books.

98 Tufekci Z, Wilson C. 2012. Social media and the decision to participate in political protest: observations from Tahrir Square. *J Commun*. 62(2):363–379.

99 Spring VL, Cameron CD, Cikara M. 2018. The upside of outrage. *Trends Cogn Sci (Regul Ed)*. 22(12):1067–1069.

100 Morozov E. 2012. *The net delusion: the dark side of internet freedom*. New York: PublicAffairs.

101 Tufekci Z. 2017. *Twitter and tear gas: the power and fragility of networked protest*. New Haven (CT): Yale University Press.

102 Chenoweth E. 2022. Can nonviolent resistance survive COVID-19? *J Hum Rights*. 21(3):304–316.

103 Chenoweth E. 2020. The future of nonviolent resistance. *J Democr*. 31(3):69–84.

104 Tufekci Z. 2014. Social movements and governments in the digital age: evaluating a complex landscape. *J Int Aff*. 68(1):1–18.

105 FeldmanHall O, Sokol-Hessner P, Van Bavel JJ, Phelps EA. 2014. Fairness violations elicit greater punishment on behalf of another than for oneself. *Nat Commun*. 5(1):5306.

106 Kanakogi Y, *et al*. 2022. Third-party punishment by preverbal infants. *Nat Hum Behav*. 6(9):1234–1242.
107 Fehr E, Fischbacher U. 2004. Third-party punishment and social norms. *Evol Hum Behav*. 25(2):63–87.
108 Leshin RA, Yudkin DA, Bavel JJV, Kunkel L, Rhodes M. 2022. Parents' political ideology predicts how their children punish. *Psychol Sci*. 33(11):1894–1908.
109 Yudkin DA, Van Bavel JJ, Rhodes M. 2020. Young children police group members at personal cost. *J Exp Psychol Gen*. 149(1): 182–191.
110 Xiao E, Houser D. 2011. Punish in public. *J Public Econ*. 95(7–8): 1006–1017.
111 Barclay P. 2006. Reputational benefits for altruistic punishment. *Evol Hum Behav*. 27(5):325–344.
112 Raihani NJ, Bshary R. 2015. The reputation of punishers. *Trends Ecol Evol (Amst)*. 30(2):98–103.
113 Balafoutas L, Grechenig K, Nikiforakis N. 2014. Third-party punishment and counter-punishment in one-shot interactions. *Econ Lett*. 122(2):308–310.
114 Boyd R, Richerson PJ. 1992. Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethol Sociobiol*. 13(3):171–195.
115 Rockenbach B, Milinski M. 2006. The efficient interaction of indirect reciprocity and costly punishment. *Nature*. 444(7120): 718–723.
116 Kamei K. 2020. Group size effect and over-punishment in the case of third party enforcement of social norms. *J Econ Behav Organ*. 175:395–412.
117 Ronson J. 2016. *So you've been publicly shamed*. New York (NY): Riverhead Books.
118 Jackson JC, *et al*. 2023. Generalized morality culturally evolves as an adaptive heuristic in large social networks. *J Pers Soc Psychol*. 125(6):1207–1238. https://doi.org/10.1037/pspa0000358.
119 Axelrod R, Hamilton WD. 1981. The evolution of cooperation. *Science*. 211(4489):1390–1396.
120 Boyd R, Richerson PJ. 1988. The evolution of reciprocity in sizable groups. *J Theor Biol*. 132(3):337–356.
121 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.
122 Tosi J, Warmke B. 2016. Moral grandstanding. *Philos Public Aff*. 44(3):197–217.
123 Westra E. 2021. Virtue signaling and moral progress. *Philos Public Aff*. 49(2):156–178.
124 Fiske AP, Rai TS, Pinker S. 2014. *Virtuous violence*. Cambridge (UK): Cambridge University Press. http://ebookcentral.proquest.com/lib/nyulibrary-ebooks/detail.action?docID=3007134
125 Jordan JJ, Sommers R, Bloom P, Rand DG. 2017. Why do we hate hypocrites? Evidence for a theory of false signaling. *Psychol Sci*. 28(3):356–368.
126 Sawaoka T, Monin B. 2018. The paradox of viral outrage. *Psychol Sci*. 29(10):1665–1678.
127 Wellman ML. 2022. Black squares for black lives? Performative allyship as credibility maintenance for social mMedia influencers on Instagram. *Soc Media Soc*. 8(1):20563051221080473.
128 Fehr E, Gachter S. 2000. Cooperation and punishment in public goods experiments. *Am Econ Rev*. 90(4):980–994.
129 Fehr E, Gächter S. 2002. Altruistic punishment in humans. *Nature*. 415(6868):137–140.
130 Jordan JJ, Rand DG. 2017. Third-party punishment as a costly signal of high continuation probabilities in repeated games. *J Theor Biol*. 421:189–202.
131 Schug J, Yuki M, Maddux W. 2010. Relational mobility explains between- and within-culture differences in self-disclosure to close friends. *Psychol Sci*. 21(10):1471–1478.
132 Yuki M, Schug J, Gillath O, Adams G, Kunkel A. 2012. Relational mobility: a socioecological approach to personal relationships. *Relationship science: integrating evolutionary, neuroscience, and sociocultural approaches*. Washington (DC): American Psychological Association. p. 137–151.
133 Yuki M, Schug J. 2020. Psychological consequences of relational mobility. *Curr Opin Psychol*. 32:129–132.
134 Cho D, Kwon KH. 2015. The impacts of identity verification and disclosure of social cues on flaming in online user comments. *Comput Human Behav*. 51:363–372.
135 Kushin MJ, Kitchener K. 2009. Getting political on social network sites: Exploring online political discourse on Facebook. *First Monday*. 14(11). https://doi.org/10.5210/fm.v14i11.2645.
136 Combs DJY, Campbell G, Jackson M, Smith RH. 2010. Exploring the consequences of humiliating a moral transgressor. *Basic Appl Soc Psych*. 32(2):128–143.
137 Klein DC. 1991. The humiliation dynamic: an overview. *J Prim Prev*. 12(2):93–121. https://doi.org/10.1007/BF02015214.
138 Globig LK, Holtz N, Sharot T. 2023. Changing the incentive structure of social media platforms to halt the spread of misinformation. *eLife*. 12:e85767.
139 Robertson CE, Del Rosario K, Rathje S, Van Bavel JJ. in press. Changing the incentive structure of social media may reduce online proxy failure and proliferation of negativity. *Brain and Behavioral Sciences*. 47(e81). https://doi.org/10.1017/S0140525X23002935.
140 Fisher M. 2022. *The chaos machine: the inside story of how social media rewired our minds and our world*. New York: Little, Brown and Company.
141 Beknazar-Yuzbashev G, Jiménez Durán R, McCrosky J, Stalinski M. 2022. Toxic content and user engagement on social media: evidence from a field experiment. *SSRN Electronic J*. https://doi.org/10.2139/ssrn.4307346. Preprint, not peer reviewed.
142 Beyens I, Loes Pouwels J, van Driel II, Keijsers L, Valkenburg PM. 2020. The effect of social media on well-being differs from adolescent to adolescent. *Scientific Reports*. 10(1). https://doi.org/10.1038/s41598-020-67727-7.
143 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):389–396.
144 Philips C. 2016. Tracking hate groups online | Welcome to Leith. Independent Lens [accessed 2023 May 19]. https://www.pbs.org/independentlens/blog/who-is-watching-the-hate-tracking-hate-groups-online-and-beyond/
145 Times TNY. 2021. Inside the capitol riot: an exclusive video investigation. The New York Times [accessed 2022 Sep 30]. https://www.nytimes.com/2021/06/30/us/jan-6-capitol-attack-takeaways.html/
146 Leatherby L. *et al*. 2021. How a presidential rally turned into a Capitol rampage. The New York Times, 12.
147 Allcott H, Braghieri L, Eichmeyer S, Gentzkow M. 2020. The welfare effects of social Media. *Am Econ Rev*. 110(3): 629–676.
148 Asimovic N, Nagler J, Bonneau R, Tucker JA. 2021. Testing the effects of Facebook usage in an ethnically polarized setting. *Proc Natl Acad Sci USA*. 118(25):e2022819118.
149 Stoyanovich J, Van Bavel JJ, West TV. 2020. The imperative of interpretable machines. *Nat Mach Intell*. 2(4):197–199.
150 Bak-Coleman JB, *et al*. 2021. Stewardship of global collective behavior. *Proc Natl Acad Sci USA*. 118(27):e2025764118.

151 Robertson CE, *et al.* 2023. Negativity drives online news consumption. *Nat Hum Behav*. 7(5):812–822.

152 Mayer RE, Fiorella L, Stull A. 2020. Five ways to increase the effectiveness of instructional video. *Educ Technol Res Dev*. 68(3): 837–852.

153 Kaufman AB, Kaufman JC. 2019. *Pseudoscience: the conspiracy against science*. Reprint edition. Cambridge (MA): MIT Press.

154 Glenn AD. 2015. Using online health communication to manage chronic sorrow: mothers of children with rare diseases speak. *J Pediatr Nurs*. 30(1):17–24.

155 Lasker JN, Sogolow ED, Sharim RR. 2005. The role of an online community for people with a rare disease: content analysis of messages posted on a primary biliary cirrhosis mailinglist. *J Med Internet Res*. 7(1):e10.

156 Lazer DMJ, *et al.* 2018. The science of fake news. *Science*. 359(6380):1094–1096.

157 Pennycook G, Rand DG. 2021. The psychology of fake news. *Trends Cogn Sci (Regul Ed)*. 25(5):388–402.

158 Bursztyn L, Handel BR, Jimenez R, Roth C. 2023. When product markets become collective traps: the case of social media (Working Paper 31771). National Bureau of Economic Research [accessed 2024 Apr 8].

159 Boyd R, Richerson PJ, Henrich J. 2011. The cultural niche: why social learning is essential for human adaptation. *Proc Natl Acad Sci USA*. 108(Supplement_2):10918–10925.