Is online moral outrage outrageous? Rethinking the indignation machine

Abstract: Moral outrage is often characterized as a corrosive emotion, but it can also inspire collective action. In this article we aim to deepen our understanding of the dual nature of online moral outrage which divides people and contributes to inclusivist moral reform. We argue that the specifics of violating different types of moral norms will influence the effects of moral outrage: moral outrage against violating harm-based norms is less antagonistic than moral outrage against violating loyalty and purity/identity norms. We identify which features of social media platforms shape our moral lives. Connectivity, omniculturalism, online exposure, increased group identification and fostering what we call “expressionist experiences”, all change how moral outrage is expressed in the digital realm. Finally, we propose changing the design of social media platforms and raise the issue of moral disillusion when ample moral protest in the online environment does not have the expected effects on the offline world.

Key words: moral outrage, social media platforms, moral psychology, collective action, polarization.

Introduction

Online moral outrage can cause disproportionately revengeful behavior toward transgressors (Gummerum, et al. 2016; Crockett 2017; Silva 2021). Justine Sacco tweeted a comment about AIDS in Africa that many considered racist. Within hours, she became the top trending topic on Twitter, millions of strangers around the world contributing to a global shaming campaign. A Minnesota dentist named Walter Palmer killed a lion in a trophy hunt in Zimbabwe. After he presented the trophy on social media, he and his family were harassed and received death threats. When talking about the #MeToo movement, Martha Nussbaum does not shy away from painting a bleak picture of those who seek mob justice on social media: “Instead of a prophetic vision of justice and reconciliation, these women prefer an apocalyptic vision in which the former oppressor is brought low, and this vision parades as justice.” (Chotiner 2021). It is said that online moral outrage not only vilifies individuals for actions that sometimes are the result of imprudent show offs, it can also exacerbate social conflict by amplifying negative attitudes toward political outgroups (Crockett 2017; Carpenter et al. 2020).

In contrast with such critical reactions, Spring et al. (2018) draw attention to the positive effects of outrage which are generally overlooked. They explore how moral outrage can inspire collective action against unjust policies, promoting the belief that participating in collective action is normatively required. For example, outrage about an ongoing conflict predicts support for nonviolent peacemaking policies (Tagar et al. 2011). Also, women who exhibit anger against men’s hostile sexist beliefs are more willing to participate in collective action for equal salaries (Becker and Wright 2011). Furthermore, recent work has explored how outrage is instrumental in making progress towards racial justice (Cherry 2021).

We agree with Spring et al. (2018; 2019) that even if outrage sometimes has negative consequences, we should not neglect its potential for positive moral impact. We also agree with Carpenter et al. (2020; see also Brady & Crockett 2019) that online moral outrage can exacerbate social conflicts. But the conditions for expressing moral outrage in the online realm are very different from those under which our disposition to be morally outraged developed through cultural evolution. So, Spring et al. (2018; 2019) underestimate how social media platforms can misdirect moral outrage. They concentrate on the psychology of intergroup relations, leaving aside the complications of digital dynamics. On the other hand, Carpenter et al. (2020) underestimate the beneficial effects of social media platforms on the function of moral outrage because they focus too much on the American culture war politics, an already highly polarized environment that undermines efforts for broad collective action.

The question of whether online outrage has, on balance, more downsides than upsides for collective action and public debates is too complex to answer decisively. Moreover, the negative effects on society might differ based on the general theories of democracy and the public sphere (i.e., the liberal, communitarian, agonistic and deliberative traditions) (Ferree et al. 2002; Althaus 2012; Wessler et al. 2021). This is why we don't take sides. In this article we aim to deepen our understanding of the conditions in which online moral outrage divides people, and in which conditions it becomes a moral force for collective action. Divisive attitudes express contempt and disdain for rival groups, distorting factual claims made by each side and amplifying mutual suspicion. Democratic values demand that citizens make decisions about public policies through inclusive discussion rather than violence, bullying, and silencing (Anderson 2006; 2022). Also, collective action is not always good for democratic societies. Undemocratic forces can form collective action as well. Recently a mob attacked the Capitol in an attempt to overturn electoral results. In what follows we focus on collective action that is compatible with egalitarian values and contributes to inclusivist moral reform.

To better understand when moral outrage facilitates collective struggles against injustices, deters moral transgressions and contributes to democratic discussion, we need to merge the literature on the evolutionary and psychological mechanisms of moral outrage with the literature on how the internet and social media shape our moral lives. Our approach has two advantages. Firstly, it does not reduce a complex social phenomenon to personal vices, such as vanity (Tosi & Warmke 2016; Nguyen & Williams 2020). According to this view, people express outrage online for self-promotion purposes, in an attempt to persuade others that they are worthy of admiration. By contrast, we show that the proliferation of moral outrage online can be explained by the interaction between our evolved mechanism for third-party punishment and the low costs for expressing outrage in the online realm. Secondly, the current debate assumes that moral outrage is a general category. We use recent psychological research on pluralist moral foundations (Haidt & Joseph 2004; Graham et al. 2013; Graham et al. 2018) to question the idea of a single definitive effect of expressing moral outrage. Our interdisciplinary approach reveals a plurality of moral experiences involved in moral outrage expressions. The content of these moral experiences leads to different consequences defined as either functional or dysfunctional. We argue that the specifics of violating different types of moral norms influences the effects of moral outrage. We suggest that moral outrage against violating harm-based norms is less antagonistic than moral outrage against violating purity and identity norms. On social media platforms one can encounter all types of moral norms violations. Because these mediums lower the costs of expressing moral outrage, foster group-based identification and are designed to keep users engaged as much as possible, they can tip the balance towards the negative effects of moral outrage expression.

In the first section, we use moral psychology and an evolutionary perspective to argue that our naturally evolved disposition for moral outrage is essentially an instrument to facilitate cooperation that responds to different types of moral norms. Protests to moral transgressions are part of our complex third-party punishment mechanism. In the second section, we identify which features of the internet and social media platforms shape how we perceive and engage in moral outrage. The way in which moral outrage is expressed in the digital realm is affected by a culture of connectivity, omniculturalism, online exposure, increased group identification and fostering what we call “expressionist experiences”. In the third section, we draw implications for evaluating the effects of online moral outrage on cooperation and public moral discourse and propose changing the design of social media platforms so that moral outrage becomes less socially corrosive. In the last section, we indicate the need for future research about how online moral protest can generate a potential novel risk for democratic civic engagement. When ample moral protest in the online environment does not have the expected effects on the offline realm, there is a serious danger of inducing moral disillusion among large mases.

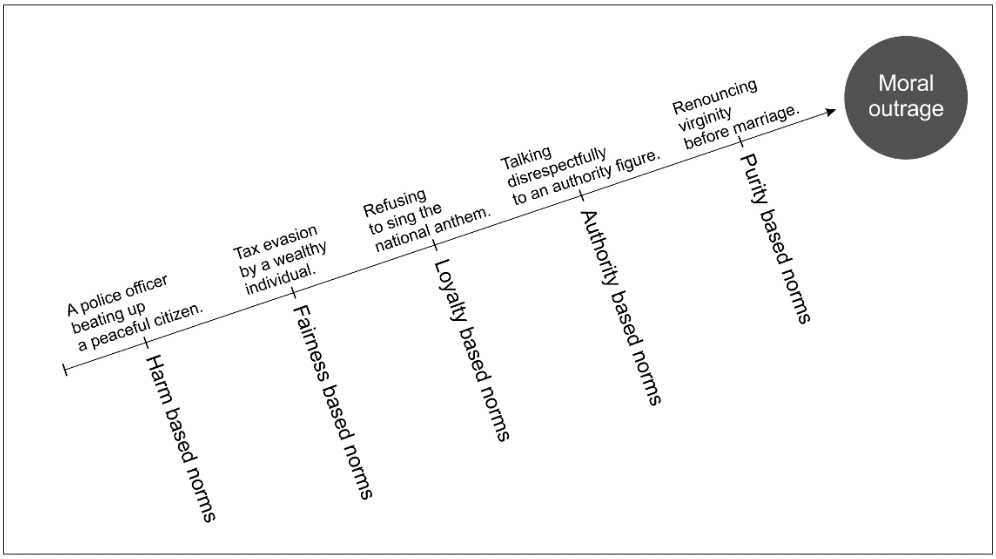
Moral outrage and cooperation: developing a pluralistic framework of the violation of norms

In this section we analyze how moral outrage manifests offline and show that it has evolved as a genuine reaction to facilitate complex cooperation, responding to different kinds of norm violations. We are a hyper cooperative species and for this reason social information is immensely valuable. The information we get from observing and communicating with others is a reliable resource for planning our actions. When we perceive that fellow human beings violate a moral norm, we react with moral outrage (Crockett 2017). The outrage mechanism is not just about noticing norm violations. One can notice the violation of a moral standard, but decide to ignore it. Thus, moral outrage is a behavioral anger response to norm violations that *translates* into shaming the violators or demanding the violators to evaluate their actions as wrongful (Srinivasar 2018; Silva 2021). A pluralistic framework of understanding the experiences of moral outrage will illustrate that taking into account differences in the content of moral norms can help to reduce socially corrosive effects.

Types of moral norms

Philosophical approaches of the moral domain have traditionally focused on one basic foundation like autonomy, avoiding harm or fairness (Kant, Mill, Rawls). Monist approaches accept that there are different moral practices across cultures, but seek a universal structure nonetheless. However, recent psychology research documents a plurality of basic moral values that are specified into different norms (Haidt and Joseph 2004; Graham et al. 2013; Graham et al. 2018). So, monist approaches can fail to appreciate how the diversity of moral norms creates different social dynamics.

Figure 1: Illustrations of different types of moral norm violations.


A police officer beating up a peaceful citizen.
Harm based norms
Tax evasion by a wealthy individual.
Fairness based norms
Refusing to sing the national anthem.
Loyalty based norms
Talking disrespectfully to an authority figure.
Authority based norms
Renouncing virginity before marriage.
Purity based norms
Moral outrage

According to the pluralist theory of moral foundations (Graham et al. 2013; Graham et al. 2018), moral norms are classified into: (1) norms against suffering (care/harm), (2) fairness norms (fairness/cheating), (3) group loyalty norms (loyalty/betrayal), (4) deference to authority and tradition norms (authority/subversion), and (5) purity norms (sanctity/degradation). Suffering, distress or neediness trigger the care/harm foundation, generating compassion for victims, which is often mixed with anger toward those who cause harm. Acts of cheating and free riding trigger the fairness foundation. People react with emotions that compel them to play “tit for tat”. The loyalty/betrayal foundation facilitates individuals to form cohesive coalitions in conditions of intergroup competition. A major consequence of betrayal is the exclusion of disloyal individuals. The authority/subversion foundation grants legitimacy to law courts, police departments, and to leaders of many kinds. Obedience and deference become virtues in the context of recognizing authority. The emotion of disgust is specific to the sanctity/degradation foundation, influencing cultural practices of treating human bodies as temples.

Expressing moral outrage is not a monolithic phenomenon. People will manifest moral outrage in response to violations of different kinds of moral norms. The plurality of experiences does not mean that there are different emotions of moral outrage, but rather that different eliciting conditions trigger one emotional mechanism. Witnessing political violence by states against their own citizens elicits moral outrage against violating a harm-based norm. Austerity measures following the economic and financial crisis that began in 2008 triggered moral outrage against violating fairness norms. Famous singer Miley Cyrus faced backlash on Twitter after revealing that she will no longer be vegan due to health issues. Her decision was considered a betrayal to the vegan movement and a threat to the reputation of the group. Authority violations consist of disrespecting authority figures (e.g., older people, teachers or parents) or symbols of authority (e.g., the courthouse, national flag). Violations of purity norms include sexually deviant acts (promiscuity, incest), allowing mases of immigrants, as well as behaviors that are seen as eroding sacred values. In general, sacred values are considered non-negotiable, as they are insensitive to material incentives and punitively rigid (Tetlock 2003). Individuals who strongly endorse sacred values show a greater willingness to fight to death (Pretus et al. 2018).

Third-party punishment facilitates cooperation

If you are personally affected by an unfair treatment, you are motivated to seek reparation through punishment. People whose economic payoff is reduced by the violation of the norm punish the violation much more strongly than do third parties (Fehr & Fischbacher 2004). When you are a non-affected party who wants to shame others for moral transgressions, the costs of moral outrage are high in day-by-day social settings. You do not have direct benefits and you can lose a lot because violators can retaliate. So, it is risky to protest when you are not the victim.

But despite what a rational calculus says, people are still willing to enforce fairness norms although they are not directly involved and it is costly for them (Fehr & Fischbacher 2004). Human beings evolved to speak out against cruelty and injustice. Third-party protest is a uniquely human ability that helps maintain cooperation by deterring free-riding and cheating. By contrast, chimpanzees have no third-party punishment (Riedl et al. 2012). They punish individuals who steal their food, but not when others' food is stolen (Riedl et al. 2012).

Human cooperation, even in large groups of genetically unrelated strangers, depends upon the enforcement of social norms and this enforcement depends on third-party punishment. What explains in part the high levels of cooperation among humans is the willingness to accuse and punish the violation of cooperation norms (Lergetporer et al. 2014). If moral outrage did not evolve, our cooperation would look more individualistic and opportunistic, similar to chimpanzee-like interactions (Rekers et al. 2011).

The social information of moral outrage facilitates cooperation

When you express moral outrage, you do not merely signal the violation of a moral standard, you also send social information about what that means to your social network and to unrelated strangers. Recent talk of virtue signaling portrays it to be sourced in vanity, without a deeper commitment to moral behavior (Tosi & Warmke 2016; Nguyen & Williams 2020). However, if people want to signal a preference for cooperation, they have to show that they are trustworthy. To do this, individuals need to endure costs, such as the costs of punishment and shaming (Henrich & Henrich 2007; Henrich 2016). Thus, third-party punishers are trusted more, and behave in a more trustworthy way than non-punishers (Jordan et al. 2016). Consequently, the third party's moral disapproval of the violation of a norm signals that the third party is unlikely to commit the same transgression and reinforces their adherence to cooperation norms. We have cross-cultural evidence that costly punishment positively covaries with altruistic behavior (Henrich et al. 2006). When a third party signals his commitment to norm adherence, they also signal a commitment to network membership, creating group cohesion (Spring et al. 2018). Notice that the process of social signaling should be viewed from a third person perspective. What others see in one's behavior becomes critical, rather than what a person intends to do. People face the challenge of extracting social information from others' actions when they don't have reliable access to their intentions and plans. So, costly actions are an indication of people's underlying true commitments and constitute credibility enhancing displays (CREDS)

(Henrich 2020). For example, one of the highest credibility enhancing display is dying for your religion. If you decide to become a martyr, then people will believe that you’re truly religious. Evolution has favored our tendency to rely on CREDS as a means against individuals who want to exploit our willingness to cooperate (Henrich 2020). It would be unlikely for you to shame and punish transgressors if you would not be committed to cooperation and norm adherence.

We argued that moral outrage is a credibility enhancing display that has evolved to facilitate higher levels of cooperation through the mechanism of third-party punishment when direct punishment, for example, fails to do so. Human cooperation depends upon the continuous enforcement of norms. What moral outrage does is to socially monitor this enforcement and trigger reactions that discourage people to violate moral norms.

In the process of deterring defection, moral outrage also signals information that provides opportunities for cooperation. Moral disapproval of norm violations signals trustworthiness and reinforces adherence to the public moral code. The signal of commitment to norm adherence is also a signal of commitment to social network membership. As such, moral outrage is not a distortion of the function of morality. On the contrary, it is a function of morality aimed at facilitating complex cooperation (Levy 2021). There is proliferation of moral outrage online and many are tempted to see this as a self-centered moral show-off. As we show later, the specifics of digital platforms explain the increase of moral outrage in the online environment.

Moral outrage on social media

Imagine someone who is scrolling on a social media website and comes across an old friend who brags about shooting one of the biggest bears in the area, legally but just for fun. As a long time, bear-lover, she is completely appalled and decides to mobilize people to physically follow, attack and harass the hunter and their family. This kind of abuse is extreme, justifying police intervention. But if the surveillance, harassment and attacks happen online, it wouldn’t alarm many people, much less the police. Similar behaviors take on different moral meanings depending on the context. In this section we assess the relevant features of the internet and of social media platforms which shape the ways in which we engage in moral outrage online.

Connectivity

The internet affords the possibility for just about everyone with a device and network connection to express and inform themselves. However, an online community is not the exclusive product of human sociality. Social media platforms engineer our sociality towards more connections. Web 2.0 gave rise to a culture of connectivity, wherein users are both recipients and consumers, producers and participants of culture (van Dijck 2012, 2013). Firstly, within digital ecosystems, sociality is co-produced between humans and machines. Users adapt and respond to the technological constraints, monetizing strategies or business strategies of social media platforms (Jacobsen 2021). The way users interact online, the things they value and the way they express themselves are mostly influenced by algorithms and the technological features of the platforms. This means that online sociality is shaped by coded structures that alter the nature of our relationships and the values we attach to them – for example, on social media platforms, ‘sharing’, ‘liking’ or ‘following’ have become social values in themselves. Secondly, platforms actively push users towards more connections. Connections mean more data, and more data means more profit. Persuasive design compels users towards sharing content with as many people as possible and encourages them to join groups that appear to be of interest to them.

Omniculturalism and exposure to moral experiences

The culture of digital connectivity creates an omnicultural medium that has implications for inter-group relations (Moghaddam 2012). Groups of people from different ethnic, religious or linguistic backgrounds connect and share experiences and knowledge. The decentralized nature of this new technology leads to a rapid flourishing of non-territorial communities that are bound by various value commitments, for example, communities of white hat hackers, vegans, feminists, young mothers, bikers, etc. (Johnson & Post 1997). Online it is much easier to join a preferred group than to try to adjust to the preferences and rules of the group we already belong to. People search for and join online communities who share their beliefs and worldviews. This homophily can strengthen the illusion that the values and principles binding the group are 'sacred' and their transgression is unpardonable (Brady et al. 2017). At the same time, the omnicultural nature of the internet makes networks, with radically different norms and *mores*, more easily visible to each other (Marwick & Boyd 2011).

The constant push for more connections and the plurality of morally salient information on social media platforms leads to a growth of information about norm violations we encounter online (Crockett 2017; Hofmann et al. 2014). This high density of moral information might explain why social media triggers such strong moral conflicts between groups that lead to moral outrage (Carpenter et al. 2020). Lowered costs of expressing outrage on social media platforms also explain the high density of morally relevant stimuli. Firstly, the costs of broadcasting, spreading and receiving knowledge of a moral norm violation are close to zero (Spring et al. 2019). Generally, expressing material moral outrage is costly because it exposes individuals to retaliation (Brady et al. 2020). Online there are no retaliation costs. People can write a comment or share a post and log-out, putting moral dialogue at arm's length. Thus, by lowering expression and retaliation costs, social media platforms foster moral outrage.

Increased group identification

Associating with similar others has come to dominate online dynamics (Cinelli et al. 2021). Compared to face-to-face communication, online communication eliminates the social cues and signals that inform how we behave towards others and respond to them. So, homophily and the depersonalized nature of online communication increase the high salience of group relations on social media, because “a specific group identity is the main relation among our social network rather than an intimate interpersonal relation” (Brady et al.2020). In and of itself, increased group identification motivates people to protect their group image (Johnen et al. 2018) which could generally yield desirable consequences – such as the mobilization of collective action (Spring et al. 2019) towards enforcing social norms that might lead to moral progress (Westra 2021). But there is evidence that when individuals strongly identify with their social groups, they are prone to dehumanize socially distal others (Waytz & Epley 2012). Moreover, when we are psychologically close to others who behave unethically towards other groups, we ourselves are inclined to behave similarly (Gino & Galinsky 2012). This might intensify the toxic expression of moral outrage online.

Expressionist experiences

Social media platforms are designed to keep us engaged as much as possible. For this purpose, most platforms use persuasive design (Williams, 2018) to exploit people’s biases. We wouldn’t normally spend so much time scrolling through never-ending newsfeeds. Statistical methods and algorithmic techniques are applied on huge databases of users’ personal data in order to extract information about their attributes and characteristics. Profiling algorithms present users with content that is most likely to elicit their emotions (Zuboff 2019). Unsurprisingly, scandalous content elicits an arousal response, which in the end makes them more likely to engage with that content. Ultimately, social media algorithms promote the spread of outrageous content.

A parallel with Impressionism and Expressionism, two artistic movements of the late 19th century, can help us distinguish between two modes of our online existence.

Impressionist painters used the artistic and scientific knowledge of light to draw serene and balanced visual scenes that invite the viewer to immerse and find their own place and peace within them. On the other hand, Expressionist painters used vivid colors to stir emotional subjective responses and to elicit intense and instant reactions, such as anxiety, in the viewer. Seeing funny pictures of domestic grumpy cats online, reading about scientific breakthroughs or morally praiseworthy people are some examples of 'Impressionist' experiences. Unfortunately, social media is not mostly about that. The moral life of information on social media is steered towards 'Expressionist' experiences, as platforms reward social status seeking, with no care for the aftermath. In the expressionist mode of online existence, users are incentivized to put on performances that will attract quantifiable attention (likes, comments and shares). Quantifying what is attractive builds the social status in the online environment. How many likes, shares, and comments one receives determines their reputation understood as online visibility.

The prevalence of 'Expressionist' online experiences leads to 'outrage fatigue' (Crockett 2017). It arises when we deal with too many moral transgressions that require our attention and as a result, we experience exhaustion and apathy, which decreases the intensity of the emotions experienced. Overloading moral signals creates a cognitive burden that impedes deeper understanding of the phenomena we deal with (Voinea et al. 2020). This way of being online is not a bug, but a structural feature chosen and implemented by platform owners. You don't react, you're not recognized by the others, and so you are not part of the game.

Implications for assessing upsides and downsides of online moral outrage

In the previous sections we argued that moral outrage is a force for complex cooperation and we identified what features of the internet in general and of social media platforms in particular determine how we engage in moral outrage online. In this section, we draw implications for assessing the effects of online moral outrage on cooperation and public moral discourse. At the moment we cannot know decisively whether online outrage has on balance more downsides than upsides. More knowledge is needed. We indicate below several implications which can further illuminate under what conditions moral outrage is corrosive and when it can lead to inclusivist moral reforms.

Polarization and types of norms

Moral outrage tends to be negatively characterized because a lot of the debate is focused on the American cultural wars. The United States is arguably among the most polarized of advanced democracies (Mccoy & Press 2022; Stewart et al. 2020). Social media can contribute to polarization, but it does so more effectively in already highly polarized environments.

The moral experiences that we encounter online contribute differently to polarization. In polarized environments almost everything can potentially get moralized and, so, it is extremely difficult for people from divergent groups to agree. Nevertheless, the moral experiences of reacting to the violations of norms are not the same because each kind of norm creates different social dynamics. For example, cheating and free riding triggers people to react with emotions that compel them to play “tit for tat”, whereas violating loyalty and purity norms triggers people to react with emotions that compel them to exclude individuals (Graham et al. 2012; Graham et al. 2018). We have to ask what kind of norms, if violated, will tend to divide and what kind of norms, if violated, will tend to bring people together. Consider the following results from the moral psychology of liberals and conservatives which suggest that people with different values can converge on harm/care norms and diverge on loyalty, authority and purity norms. Liberals show greater endorsement of the Harm/care and Fairness/reciprocity norms, whereas conservatives endorse more equally the Harm/care, Fairness/reciprocity, Ingroup/loyalty, Authority/respect, and Purity/sanctity norms (Graham et al.2009). Liberals endorse the moral concerns of compassion and fairness more than conservatives do, and conservatives endorse the moral concerns of ingroup loyalty, respect for authorities and traditions, and physical/spiritual purity more than liberals do (Graham et al. 2012).

It seems that if the norms violated are identity-based norms (purity, sacred values, loyalty), it is likely that people will tend to divide, given that the function of identity-based norms is to prepare ingroup members for competition with other groups. We do not claim that harm violations will always make people reach common ground. We only make a comparative claim that it is more likely to reach common ground against perpetrators when harm-based norms are violated rather than when group identity norms are violated. The omnicultural nature of the internet can accentuate social conflicts because it makes online communities, with radically different values, more visible to each other. So, individuals are motivated to maintain their social status in relation to a specific group identity (Brady et al. 2020). An emotional content that is based on group-identity motivations will likely capture people's attention, and, consequently, the design of social-media platforms will foster the spread of such content (Brady et al. 2020).

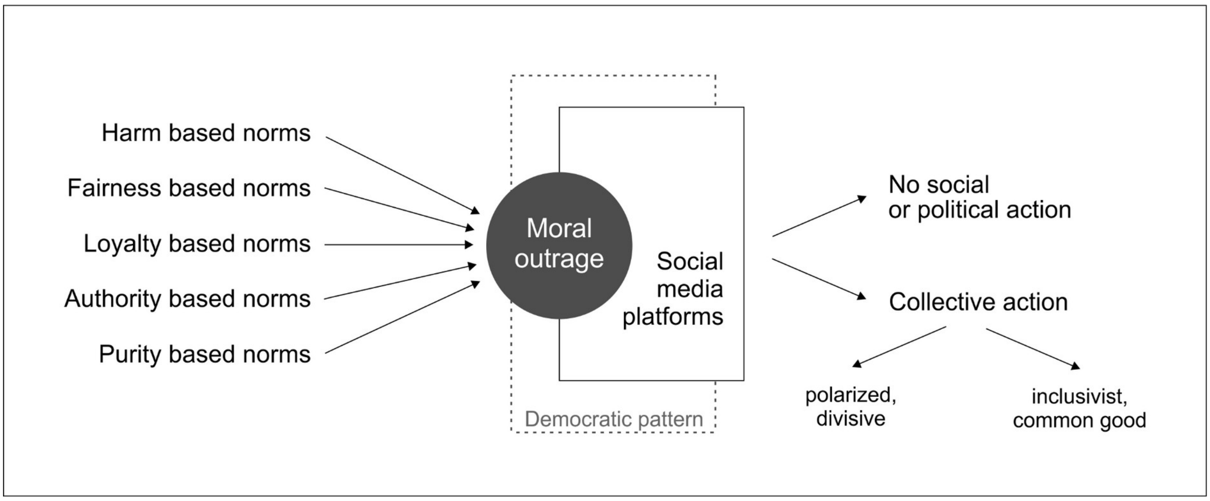
Whereas if the norms violated are harm-based, people will tend to reach common ground because harm/care norms are shared more universally (Kinnier et al. 2000) compared to loyalty and purity norms which are conditioned by a local context. Identity norms are instrumental to forming close knit groups, whereas in contexts of intergroup competition they facilitate antagonism (Appiah 2004). The moral emotions triggered by the violation of each type of norm generate different social interactions. For example, moral anger is more flexible than moral disgust. People find it difficult to imagine circumstances that potentially mitigate the moral wrongness of purity violations compared to harm violations (Russell & Giner-Sorolla 2011). Further, exploring how people deal with group conflicts can reveal that some moral foundations are punitively more rigid than other foundations. There is some evidence that individuals prefer to harm their own group rather than help an opposing group across polarized issues (abortion access, political party, gun rights) (Gershon & Fridman 2022).

We suggest that online moral outrage can broaden collective support against political violence, human rights abuses, and in general violations of harm based norms. Social media connectivity was instrumental to communication as well as to the dissemination of information for social movements in the Arab uprisings (Rane & Salem 2012). The use of social media did not precede, but rather followed a significant amount of protest activity (Wolfsfeld et al. 2013). Platforms are an important tool in scaling protests as they reduce the costs of collective action, thus helping people bypass the construction of formal structures which need time and a lot of interaction between the organizers and the participants (Tufekci 2017). They provide an important means to spread information internationally by avoiding traditional mass-media gatekeepers and to harness support. Take, for example, the current anti-government protests in Tehran. Irani citizens shared videos of their participation in protests and of the violent response from the authorities. These videos became viral internationally, motivating people from democratic countries to show solidarity (The Guardian, 2022).

It is not our aim here to discuss why some displays of solidarity on social media are short-lived or whether people are exhibiting mere ‘clicktivism’. Our point is that human rights abuses or extreme violence have less potential to polarize and are more prone to garner support for victims. A recent systematic review found that the use of social media platforms increased political participation and information consumption in autocracies and emerging democracies, while it increased populism, polarization and political distrust in established democracies (Lorenz-Spreen et al. 2022). Human rights abuses are more frequent in autocracies and emerging democracies, and so, online moral outrage is useful to counter authoritarian ambitions. Established democracies are characterized by allowing the expression of multiple differences which can increasingly be perceived in terms of “us” versus “them.” It seems that in established democracies, online moral outrage may predominantly highlight problems of transgressing identity norms.

Viral spread of human rights abuses at the global level could also mobilize people to offer humanitarian aid by generating a sense of expanding solidarity. How the Russian invasion of Ukraine faced immediate condemnation from all around the world offers plausibility to this hypothesis. The war has generated an ongoing humanitarian and refugee crisis in Ukraine, but the use of social media has stimulated the expression of solidarity and garnered support for refugees (Zawadzka-Paluektau 2022; De Coninck 2022).

Figure 2: Illustration of the mechanisms and effects of online moral outrage.



Reforming social media

Online we are removed from the suffering we inflict on those we punish through our moral outrage, which of course makes the infliction of harm much easier (Cocking & Van den Hoven 2018). In fact, we are not physically harming others, we are just commenting and sharing, which seem morally benign actions. But when a mere comment is part of a "cascade" of similar vitriolic comments, the consequences can be devastating. Online evil is cumulative. Seemingly harmless actions (commenting or sharing) done simultaneously by thousands or even hundreds of thousands of people create disproportionately punitive environments (Cocking & Van den Hoven 2018).

If the main purpose of social media platforms is to maximize how long users are active online, then they will implicitly facilitate ‘outrage cascades’ that oftentimes degenerate in bullying, public shaming and unjust or even cruel behaviors. These phenomena keep people engaged and scrolling for longer (Fritz 2021). Currently, platforms have no incentives to stop the continuous formation of online ‘outrage cascades’, and neither is it in their interests to provide more ‘impressionist’ content to users.

There is no technological necessity for social media platforms to look as they do. The current design of social media is arbitrary. Changing the attention-driven data economy could make online moral outrage strive for inclusivist moral reform The online environments where we spend increasing amounts of time are by no means neutral instruments for sharing and exchanging information. Can we reform social media so that it no longer speculates human psychology for an attention seeking economy?

Undoubtedly, yes. Social media platforms should change the targeted delivery of messages and ads based on users’ personal data. Algorithms now work to increase engagement with the price of spreading morally inflammatory content, mis- and disinformation or other types of morally problematic content (Williams 2018; Benkler, Robert, and Hal 2018; Brady et al. 2020). Instead of promoting intensive engagement, algorithms could be designed to enhance self-control, emotion recognition, undivided attention and responsibility beliefs.

Firstly, algorithms could slow down users' impulses, ‘nudging’ them to reconsider their immediate reactions. We nudge someone when we arrange her choice context in order to influence the likelihood of choosing option A over option B, even though it would still be easy to choose B. Nudging interventions have been successfully implemented in many areas of public policies that address obesity, smoking, distracted driving, food safety, organ donation (Thaler & Sunstein 2008; Sunstein 2014; Mihailov 2019). Reforming the ways social media engages users could benefit from the extensive research on nudging interventions (Thornhill et al. 2019). Before ‘posting’, ‘linking’ or ‘sharing’, for example, a pop-up message could ask users if they are sure about what they are planning to do. Nudges presented alongside relevant information can make people self-conscious regarding the epistemic status of their beliefs. We can improve the current design of social media platforms through user and political pressure that can be fed, ironically, also by moral outrage.

Secondly, we should design social media to become less status oriented (less expressionistic). In the early days of social media platforms, users shared information and experiences to close ones and professional peers. Now, algorithms incentivize users to frame their online activity in terms of building personal brands and enhance their visibility (Haidt 2022). We suggest changing the quantifiable approach to social status and approval. Facebook, Twitter and Instagram could reduce the extent to which online content is publicly quantified. One important consequence is that the number of likes and shares one receives for creating content will not be visible to others. The user can see the total numbers of online engagement, but they will not be available to the other members of the social network. Facebook and Instagram have implemented this feature but only as an option, not as the default. We should strive to do more of the same.

Baiting crowds

Crowds are not always engaging in collective actions for changes towards a greater good, the restoration of justice, or for inclusivist moral reforms. A significant risk that we need to take into account is as old as the first social organizations: the formation of “the baiting crowds”, thirsty for blood and justice, in need to find a scapegoat. Forming the baiting, aggressive crowd only needs a clear target, a goal to follow (Canetti 1973, 49). The public execution, Canetti (1973, 50) said, is reminiscent of the “old practice of collective killing”. Of course, we civilized people no longer witness public executions. Indeed, we now reject public displays of physical violence, but we still take part, through the media, in public displays of moral violence. Social media platforms are the perfect locus for baiting crowd formation, in a more effective form than in older media. Canetti saw, more than 50 years ago, the newspaper as an exemplary informational medium through which we participate in public executions from a distance. Social media provides new affordances for public executions, and this is the result of a culture of permanent connectivity that exploits human connectedness by commodify it (van Dijck 2013, 16). For some years now, many platforms have opened the stage to thrilling ‘Expressionism’. Anger is demanded and consumed, generating collective moral chaos in circumstances of ambiguity, conflictual normative communities, and persuasive technologies. What is at stake here is to inquire into how social media platforms blur the boundaries between in-groups and out-groups and how they could incentivize people to show solidarity beyond their immediate moral communities.

Moral disillusion: the need for future research

To learn more about how online moral outrage can shape an expanding sense of solidarity we need further research about when online moral outrage correlates with actual helping behavior or donations. For example, online moral outrage has translated into helping behavior during the humanitarian crisis generated by the Russian war against Ukraine. Airbnb, one of the largest online platforms for renting, announced that it will help with housing for refugees. Online communities have been mobilized through social media to support Airbnb’s initiatives towards Ukraine, attracting millions of supporters (Cheng 2022). The success of the Airbnb initiative lies in its effective use of social media to mobilize collective action.

Effective use of online moral outrage depends on many things. Further research should explore how and whether online moral outrage translates into actual protest and costly collective behavior. Remember that expressing moral outrage in the offline world exposes individuals to retaliation, a cost which signals that individuals are trustworthy. However, the increase of online moral outrage due to lower costs should be interpreted with care. Expressing outrage in online conditions no longer indicates reliably people’s willingness to participate in costly collective action. Its characteristic low costs, as opposed to costly moral outrage in the offline world, questions the status of credibility enhancing display of online moral outrage. This makes us vulnerable to exploitative outrage which incites negative sentiment solely for the purpose, for example, of gaining political advantage. If online moral outrage rarely translates into collective action, we should start talking about moral disillusion as a factor that contributes to people’s disengagement from democratic and civic processes. It would be a dissonant world if we see that moral protest is boiling online, without spilling over into the offline world. The translation problem could make people lose faith in solving public moral issues, reinforcing existing social inequalities and providing opportunities for authoritarian movements. Widespread online moral outrage could create the false expectation that reality will change to accommodate people’s demands, whereas changing the *status quo* requires so much more. People’s frustration will be there to be exploited.

We need to explore what happens to people’s social perceptions if constant and ample moral protest in the online environment does not have the expected effects on the offline world. If the online anger and outrage of marginalized communities are more difficult to translate into effective social action, then this reinforces the marginalization of those communities. In contexts of political engagement, the potential of online moral outrage to harness collective action depends on the societal limits of outrage expression (Phoenix 2020). Some groups are socially permitted to express anger, outrage and grievance, while others are not. There is significant evidence that outrage motivates some demographics and not others in the context of politics (Phoenix 2020). Democratic engagement risks deteriorating if social media platforms facilitate the expression of moral outrage but the social inequalities and social limits remain in place to block collective action.

Conclusion

Moral outrage fosters complex cooperation. It should not worry us too much in itself. When assessing the social effects of online moral outrage, we should take into consideration what types of norms are violated. Transgressing identity-based norms has greater potential to polarize people than in the case of harm-based norms, where we see greater consensus on the importance of norms. There is a proliferation of moral outrage cascades on social media platforms because online punishments are apparently harmless and the costs of expressing outrage online are lower compared to offline. Thus, we should interpret with care the social signal of online moral outrage. Further research should explore whether and how online moral outrage correlates with people's willingness to endure altruistic costs. Moreover, social media platforms offer the perfect settings for the formation and manifestation of baiting crowds. What is at stake is how to reform current social media platforms to offer 'Impressionist' modes of existence, digital spaces for democratic deliberation and channels for mobilizing inclusivist collective action.

Acknowledgements

We are grateful to Alexandra Zorilă, Mihaela Constantinescu, Anda Zahiu, Cristian Iftode, Veronica Lazăr, Bettina Lange, and Dan Zeman for thoughtful comments. We also thank the editor and reviewers for their advice and comments throughout the reviewing process.

Funding

This work was supported by a grant of the Romanian Ministery of Education and Research, CNCS – UEFISCDI, project number PN-III-P4-ID-PCE-2020–0521, within PNCDI III.

References

Appiah, K. A. (2004). *The Ethics of Identity*. Princeton University Press.

Becker JC, and Wright SC (2011) Yet another dark side of chivalry: Benevolent sexism undermines and hostile sexism motivates collective action for social change. *Journal of personality and social psychology*, 101(1): 62-77.

Brady WJ, and Crockett MJ (2019) How effective is online outrage? *Trends in cognitive sciences*, 23(2): 79-80.

Brady WJ, Crockett MJ, and Van Bavel JJ (2020) The MAD Model of Moral Contagion: The Role of Motivation, Attention, and Design in the Spread of Moralized Content Online. *Perspectives on Psychological Science* 15 (4): 978–1010. https://doi.org/10.1177/1745691620917336.

Brady WJ, Wills JA, Jost JT, Tucker JA, & Van Bavel JJ (2017) Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114 (28): 7313-7318.

Brady, W. J., Crockett, M. J., & Van Bavel, J. J. (2020). The MAD model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science*, 15(4), 978-1010.

Canetti E (1981) *Crowds and Power*. New York: Continuum.

Carpenter J, William BJ, Crockett MJ, Weber R, and Sinnott-Armstrong W (2020) Political Polarization and Moral Outrage on Social Media. *Connecticut Law Review* 52 (3): 1107–1120.

Chalaby JK, and Plunkett S (2021) Standing on the Shoulders of Tech Giants: Media Delivery, Streaming Television and the Rise of Global Suppliers. *New Media & Society* 23 (11): 3206–28. https://doi.org/10.1177/1461444820946681.

Cheng, M. (2022). Mobilize Airbnb support in times of humanitarian crisis. *Current Issues in Tourism*, 1-7.

Cherry, M (2021) *The Case for Rage: Why Anger Is Essential to Anti-Racist Struggle*. Oxford University Press.

Chotiner I (2021) Martha Nussbaum on #MeToo. *The New Yorker*, June 1, 2021. https://www.newyorker.com/news/q-and-a/martha-nussbaum-on-metoo.

Cinelli, M., Morales, G. D. F., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021) The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9). https://doi.org/10.1073/pnas.2023301118

Cocking D and van den Hoven J (2018) *Evil Online*. New York: Wiley-Blackwell.

Crockett M (2017) Moral Outrage in the Digital Age. *Nature Human Behaviour* 1 (11): 769–71. https://doi.org/10.1038/s41562-017-0213-3.

De Coninck, D. (2022). The Refugee Paradox During Wartime in Europe: How Ukrainian and Afghan Refugees are (not) Alike. *International Migration Review*, 0(0). https://doi.org/10.1177/01979183221116874

Dobele A, Lindgreen A, Beverland M, Vanhamme J, and van Wijk R (2007) Why Pass on Viral Messages? Because They Connect Emotionally'. *Business Horizons* 50 (4): 291–304. https://doi.org/10.1016/j.bushor.2007.01.004.

Domingos P (2015) *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. New York: Basic Books.

Fehr E, and Fischbacher U (2004). Social norms and human cooperation. *Trends in cognitive sciences*, 8(4): 185-190.

Fritz J (2021) Online Shaming and the Ethics of Public Disapproval. *Journal of Applied Philosophy,* 38: 686-701 https://doi.org/10.1111/japp.12510.

Gershon, R., & Fridman, A. (2022). Individuals prefer to harm their own group rather than help an opposing group. *Proceedings of the National Academy of Sciences*, 119(49), e2215633119.

Gillespie T (2017) Governance of and by Platforms. In: Burgess J, Poell T, and Marwick A (Eds) *SAGE Handbook of Social Media*. London: SAGE Publications, 254–78.

Gino F, and Galinsky AD (2012) Vicarious Dishonesty: When Psychological Closeness Creates Distance from One's Moral Compass. *Organizational Behavior and Human Decision Processes* 119 (1): 15–26. https://doi.org/10.1016/j.obhdp.2012.03.011.

Graham J, Haidt J, and Nosek BA (2009) Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5): 1029-1046.

Graham J, Haidt J, Koleva S, Motyl M, Iyer R, Wojcik SP, and Ditto PH. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in experimental social psychology*, 47: 55-130.

Graham J, Nosek BA, and Haidt J. (2012). The moral stereotypes of liberals and conservatives: Exaggeration of differences across the political spectrum. *PloS one*, 7(12): e50092.

Graham, J., Haidt, J., Motyl, M., Meindl, P., Iskiwitch, C., & Mooijman, M. (2018). Moral foundations theory: On the Advantages of Moral Pluralism over Moral Monism. In Gray, K. & Jesse Graham (eds.), *Atlas of moral psychology*, Guilford Press.

Gummerum M, Van Dillen LF, Van Dijk E, and López-Pérez B (2016) Costly third-party interventions: The role of incidental anger and attention focus in punishment of the perpetrator and compensation of the victim. *Journal of Experimental Social Psychology* 65: 94-104.

Haidt, J. (2022). Why the Past 10 Years of American Life Have Been Uniquely Stupid. *The Atlantic*, 11.

Haidt J, and Joseph C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4): 55-66.

Haidt J, and Rose-Stockwell T. (2019). The dark psychology of social networks. The Atlantic, 6-60.

Henrich, N., & Henrich, J. P. (2007). *Why humans cooperate: A cultural and evolutionary explanation*. Oxford University Press.

Henrich, J. (2016). *The Secret of Our Success*. Princeton University Press.

Henrich, J. (2020). *The WEIRDest people in the world: How the West became psychologically peculiar and particularly prosperous*. Penguin UK.

Henrich J, McElreath R, Barr A, Ensminger J, Barrett C, Bolyanatz A, and Ziker J (2006) Costly punishment across human societies. *Science*, 312(5781): 1767-1770.

Hofmann W, Wisneski DC, Brandt MJ, and Skitka LJ (2014) Replication Data for: Morality in Everyday Life. *Harvard Dataverse*. https://doi.org/10.7910/DVN/26910.

Jacobsen BN (2021) Regimes of Recognition on Algorithmic Media. *New Media & Society*. https://doi.org/10.1177/14614448211053555.

Janssen MA, and Bushman C (2008) Evolution of cooperation and altruistic punishment when retaliation is possible. *Journal of theoretical biology*, 254(3): 541-545.

Johnen M, Jungblut M, and Ziegele M (2018) The Digital Outcry: What Incites Participation Behavior in an Online Firestorm? *New Media & Society* 20 (9): 3140–60. https://doi.org/10.1177/1461444817741883.

Johnson DR, and Post DG (1997) Law and Borders - the Rise of Law in Cyberspace. SSRN Scholarly Paper ID 535. Rochester, NY: *Social Science Research Network*. https://papers.ssrn.com/abstract=535.

Jordan J, McAuliffe K, and Rand D (2016) The effects of endowment size and strategy method on third party punishment. *Experimental Economics*, 19(4): 741-763.

Kinnier, R. T., Kernes, J. L., & Dautheribes, T. M. (2000). A short list of universal moral values. *Counseling and values*, 45(1), 4-16.

Klinger U, and Svensson J (2018) The End of Media Logics? On Algorithms and Agency. *New Media & Society* 20 (12): 4653–70. https://doi.org/10.1177/1461444818779750.

Lergetporer P, Angerer S, Glätzle-Rützler D, and Sutter M (2014) Third-party punishment increases cooperation in children through (misaligned) expectations and conditional cooperation. *Proceedings of the National Academy of Sciences*, 111(19): 6916-6921.

Levy N (2021) Virtue signalling is virtuous. *Synthese*, 198(10): 9545-9562.

Lewis R, Marwick AE, and Partin WC (2021) “We Dissect Stupidity and Respond to It”: Response Videos and Networked Harassment on YouTube. *American Behavioral Scientist* 65 (5): 735–56.

Lorenz-Spreen, P., Oswald, L., Lewandowsky, S., & Hertwig, R. (2022). A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature Human Behaviour*, 1-28.

Marlowe FW, Berbesque JC, Barr A, Barrett C, Bolyanatz A, Cardenas JC, and Tracer D (2008) More 'altruistic' punishment in larger societies. *Proceedings of the Royal Society B: Biological Sciences*, 275(1634): 587-592.

Marwick AE, and boyd d (2011) I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience. *New Media & Society* 13 (1): 114–33. https://doi.org/10.1177/1461444810365313.

Mccoy J, Press B (2022) What Happens When Democracies Become Perniciously Polarized?. *Carnegie Endowment for International Peace*. https://carnegieendowment.org/2022/01/18/what-happens-when-democracies-become-perniciously-polarized-pub-86190

Mihailov, E (2018). Refocusing the Nudge Debate on Organ Donation. In Mihailov E, Wangmo T, Federiuc V & Elger B (eds) *Contemporary debates in bioethics: European perspectives*, pp. 1-174.

Moghaddam FM (2012) The Omnicultural Imperative. *Culture & Psychology* 18 (3): 304–30. https://doi.org/10.1177/1354067X12446230.

Nguyen CT, and Williams B (2020) Moral outrage porn. *Journal of Ethics and Social Philosophy* 18(2): 147-172.

Quattrociocchi W, Scala A, and Sunstein CR (2016) Echo Chambers on Facebook. SSRN Scholarly Paper ID 2795110. Rochester, NY: *Social Science Research Network*. https://papers.ssrn.com/abstract=2795110.

Pretus, C., Hamid, N., Sheikh, H., Ginges, J., Tobeña, A., Davis, R., ... & Atran, S. (2018). Neural and behavioral correlates of sacred values and vulnerability to violent extremism. *Frontiers in psychology*, 2462.

Phoenix, D. L. (2019). *The anger gap: How race shapes emotion in politics*. Cambridge University Press.

Rane H, and Salem S (2012) Social media, social movements and the diffusion of ideas in the Arab uprisings. *Journal of international communication*, 18(1): 97-111.

Rekers, Y., Haun, D. B., & Tomasello, M (2011) Children, but not chimpanzees, prefer to collaborate. *Current Biology*, 21(20), 1756-1758.

Riedl K, Jensen K, Call J and Tomasello M (2012) No third-party punishment in chimpanzees. *Proceedings of the national academy of sciences*, 109(37): 14824-14829.

Russell, P. S., & Giner-Sorolla, R. (2011). Moral anger is more flexible than moral disgust. *Social Psychological and Personality Science*, 2(4), 360-364.

Schein C, and Gray K (2015) The unifying moral dyad: Liberals and conservatives share the same harm-based moral template. *Personality and Social Psychology Bulletin*, 41(8): 1147-1163.

Shteynberg G (2018) A collective perspective: Shared attention and the mind. *Current Opinion in Psychology*, 23: 93-97.

Silva L (2021) Is anger a hostile emotion? *Review of Philosophy and Psychology*, pp: 1-20.

Spring VL, Cameron CD, and Cikara M. (2018). The upside of outrage. *Trends in Cognitive Sciences*, 22(12): 1067-1069.

Spring VL, Cameron DC, and Cikara M (2019) Asking Different Questions about Outrage: A Reply to Brady and Crockett. *Trends in Cognitive Sciences* 23 (2): 80–82. https://doi.org/10.1016/j.tics.2018.11.006.

Srinivasan A (2018) The aptness of anger. *Journal of Political Philosophy*, 26(2): 123-144.

Stein DH, Schroeder J, Hobson NM, Gino F, and Norton MI (2021) When alterations are violations: Moral outrage and punishment in response to (even minor) alterations to rituals. *Journal of personality and social psychology*. Epub ahead of print, https://doi.org/10.1037/pspi0000352.

Stewart AJ, McCarty N, and Bryson JJ (2020) Polarization under rising inequality and economic decline. *Science advances*, 6(50): eabd4201.

Sunstein, CR (2014) *Why nudge?: The politics of libertarian paternalism*. Yale University Press.

Tagar MR, Federico CM, and Halperin E (2011) The positive effect of negative emotions in protracted conflict: The case of anger. *Journal of Experimental Social Psychology*, 47(1): 157-164.

Tetlock, P. E. (2003) Thinking the unthinkable: Sacred values and taboo cognitions. *Trends in cognitive sciences*, 7(7), 320-324.

Thaler RH, and Sunstein CR (2008) *Nudge: Improving decisions about health, wealth, and happiness*. Yale: Yale University Press.

The Guardian (2022) *Protest strikes in Iran reported as solidarity rallies held around world*, accessed 28 November 2022, https://www.theguardian.com/world/2022/oct/22/protest-strikes-in-iran-reported-as-solidarity-rallies-held-around-world

Thomas EF, Cary N, Smith LGE, Spears R, and McGarty C (2018) The Role of Social Media in Shaping Solidarity and Compassion Fade: How the Death of a Child Turned Apathy into Action but Distress Took It Away. *New Media & Society* 20 (10): 3778–98. https://doi.org/10.1177/1461444818760819.

Thornhill, C., Meeus, Q., Peperkamp, J., & Berendt, B. (2019) A digital nudge to counter confirmation bias. *Frontiers in big data*, 2, 11.

Tosi J, and Warmke B. (2016) Moral grandstanding. *Philosophy & Public Affairs*, 44 (3): 197-217.

Tufekci, Z. (2018) Twitter and Tear Gas. New Haven, CT: Yale University Press.

Van Dijck J (2012) Facebook as a tool for producing sociality and connectivity. *Television & new media*, 13(2): 160-176.

van Dijck J (2013) *The Culture of Connectivity: A Critical History of Social Media*. New York: Oxford University Press.

Voinea C, Vică C, Mihailov E, and Savulescu J (2020) The Internet as Cognitive Enhancement. *Science and Engineering Ethics* 26 (4): 2345–62. https://doi.org/10.1007/s11948-020-00210-8.

Waytz A, and Epley N (2012) Social Connection Enables Dehumanization. *Journal of Experimental Social Psychology* 48 (1): 70–76. https://doi.org/10.1016/j.jesp.2011.07.012.

Westra E (2021) Virtue Signaling and Moral Progress. *Philosophy & Public Affairs* 49 (2): 156–78. https://doi.org/10.1111/papa.12187.

Williams, J (2018) *Stand out of Our Light: Freedom and Resistance in the Attention Economy*. Cambridge: Cambridge University Press.

Wolfsfeld, G., Segev, E., & Sheafer, T. (2013). Social media and the Arab Spring: Politics comes first. *The International Journal of Press/Politics*, 18(2), 115-137.

Wu T (2017) *The Attention Merchants: The Epic Struggle to Get Inside Our Heads*. Bloomsbury: Atlantic Books.

Yochai B, Faris R, and Roberts H (2018) *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. Oxford: Oxford University Press.

Zuboff S (2019) *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. London: Profile Books.

Zawadzka-Paluektau, N (2022) Ukrainian refugees in Polish press. *Discourse & Communication*, https://doi.org/10.1177/17504813221111636